# *MYCOBACTERIUM TUBERCULOSIS INFECTION IN SOUTHERN AFRICA – EXPLORING PATTERNS, LOCATING TRANSMISSION*

*Dr Thomas Alexander Yates*

*University College London*

*Research Department of Infection and Population Health*

*PhD Thesis*

I, Thomas Alexander Yates, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

## Abstract

Tuberculosis is a major cause of premature mortality. Communities in Southern Africa are disproportionately affected. A growing body of evidence suggests that recent transmission within households can explain only a limited proportion of tuberculosis disease. However, our understanding of where transmission between households occurs is limited.

I undertook a systematic review and meta-analysis of molecular epidemiology studies that described rates of strain discordance in co-prevalent cases of tuberculosis resident in the same household. I also conducted a tuberculin school survey in 6-8 year old children in a rural community in Northern KwaZulu-Natal. These children were all registered in a household surveillance programme operated by the Africa Centre for Population Health.

I found that, across a range of both high and low burden countries, co-prevalent cases of tuberculosis in the same household often have different strains of *Mycobacterium tuberculosis.* These molecular epidemiological data suggest, at least in some settings, that recent transmission within households may explain a modest proportion of tuberculosis disease.

I estimated the annual risk of tuberculous infection to be approximately two percent in the community around the Africa Centre. I found weak evidence that exposure to HIV positive adults in the household was associated with *Mycobacterium tuberculosis* infection in children. I found no strong evidence associating use of specific indoor public spaces with *Mycobacterium tuberculosis* infection.

Transmission between households is likely an important determinant of tuberculosis disease. Further research locating *Mycobacterium tuberculosis* transmission might enable TB control interventions to be better targeted.

# Table of Contents

## List of Tables

## List of Figures

## Abbreviations

| | |
|---|---|
| ACH | Air Changes per Hour |
| aOR | Adjusted Odds Ratio |
| ART | Antiretroviral Therapy |
| ARTI | Annual Risk of Tuberculous Infection |
| BCG | Bacillus Calmette-Guérin vaccine |
| CD4 | Cluster of Differentiation 4 (a glycoprotein found on white blood cells, including the subset of T lymphocytes targeted by HIV) |
| CI | Confidence Interval |
| DNA | Deoxyribonucleic Acid |
| ELISPOT | Enzyme-Linked ImmunoSpot |
| HIV | Human Immunodeficiency Virus |
| IGRA | Interferon Gamma Release Assay |
| KZN | KwaZulu-Natal |
| LTBI | Latent (*Mycobacterium) tuberculosis* Infection |
| MDR-TB | Multidrug-Resistant Tuberculosis |
| MIRU-VNTR | Multiple Interspersed Repetitive Units – Variable Number of Tandem Repeats |
| MTB | *Mycobacterium tuberculosis* |
| NTM | Non-tuberculous Mycobacterium |
| OR | Odds Ratio |
| PCA | Principal Components Analysis |
| PCR | Polymerase Chain Reaction |
| PET-CT | Positron Emission Tomography – Computed Tomography |
| PHC | Primary Healthcare Clinic |
| PPD | Purified Protein Derivative |

| | |
|---|---|
| RDP | Reconstruction and Development Programme |
| RFLP | Restriction Fragment Length Polymorphism |
| SNP | Single Nucleotide Polymorphism |
| TB | Tuberculosis |
| TST | Tuberculin Skin Test |
| UVGI | Ultraviolet Germicidal Irradiation |
| WGS | Whole Genome Sequencing |
| WHO | World Health Organization |
| XDR-TB | Extensively Drug-Resistant Tuberculosis |

## Acknowledgements

I would first like to thank my supervisors Frank Tanser and Ibrahim Abubakar for their support and encouragement, for being enthusiastic about my ideas and for huge amounts of pragmatic advice. I would like to thank the Medical Research Council, the KwaZulu-Natal Research Institute for Tuberculosis and HIV (K-RITH) and the UCL Grand Challenges Small Grant Scheme for funding my research. I am grateful to Rob Aldridge for his friendship and diligent assistance in the more tedious aspects of undertaking the systematic review. Ngiyabonga kakhulu to many people at the Africa Centre for patiently showing me the ropes – especially Zanomsa Gqwede, Tinofa Mutevedzi, Mbusi Ngema and Colin Newell. Huge thanks to the tuberculin survey team – Siphiwe Cebekhulu, Mumsy Mthethwa and Happiness Mkhwanazi – for being great company, for their resourcefulness and for lots of hard work. I am grateful to the children, families and schools who participated in the survey and to Bonnie Davis and others at Hlabisa Hospital for seeing the children with positive reactions who needed clinical review. Thanks also to Nothando Sabela, Richard Lessells, Joerg and Janet Michel, Dickman Gareta, Collins Iwuji, Kevi Naidu, Maureen Hlongwane and Jaffer Zaidi for being good friends in a small town. I am grateful to Caroline De Brun, Kathy Baisley, Pally Khan, and the participants at the 2015 TSRU meeting for their ideas regarding the analysis. I am grateful to Derek Young for advice about his R package *mixtools*. Finally, thanks to Rosie for tolerating my protracted absences with her usual good humour.

# 1. Introduction

The research presented in this thesis is an attempt to better understand *Mycobacterium tuberculosis* (MTB) epidemiology, particularly in Southern Africa where, by any measure, tuberculosis (TB) is out of control(1,2). A particular focus of the work is my wish to better locate transmission. If we better understood where MTB transmission occurs, TB control programmes might be better targeted(3). The thesis consists of an introduction, three results chapters, and a discussion.

In this chapter, I briefly describe the tools we have for studying TB epidemiology, leaving detailed discussion of the tools used in TB molecular epidemiology and of the tuberculin skin test for Chapters 2 and 3, respectively; I outline current understanding of the natural history of TB disease and of airborne MTB transmission; I summarise what is known about risk factors for MTB infection, with a focus on exposure to putative sites of MTB transmission; and I describe TB epidemiology in KwaZulu-Natal (the province where much of this research was undertaken) and how the epidemic there compares to that in the wider region. My thoughts on approaches to achieving TB control, incorporating insights from the three results chapters, can be found in the discussion (Chapter 5).

Much of this research was undertaken at the Wellcome Trust Africa Centre for Population Health, which operates a household surveillance programme in Northern KwaZulu-Natal(4). In Chapter 3, there is a detailed description of that research platform.

## Tools for studying TB epidemiology

MTB is not an easy organism to study. The available tools for diagnosing infection and disease have poor sensitivity and or specificity. Our understanding of the organism's natural history remains incomplete. Prevalence and annual incidence of disease (and the incidence of infection in many communities) rarely exceed 2%. This means large studies are needed to make inferences about TB epidemiology at a population level(5).

### Tests for MTB infection

The World Health Organization (WHO) have adopted a pragmatic definition of latent MTB infection (LTBI) as 'a state of persistent immune response to stimulation by *Mycobacterium tuberculosis* antigens without evidence of clinically manifested active TB' (6). There are two main types of test for 'MTB infection' – the tuberculin skin test (TST) and the interferon gamma release assay (IGRA). Both look for evidence of a previous adaptive immune response to MTB.

#### *Tuberculin skin test*

The TST is probably the most widely used immunological test globally. It is the textbook example of a delayed type (or type IV) hypersensitivity reaction. The fascinating history of its development and standardisation is detailed in Hans Rieder's monograph(7).

Tuberculin is a 'purified protein derivative' (PPD) produced from cultures of MTB. The TST involves injecting a standard volume of PPD intradermally, typically into the volar aspect of the forearm, then measuring the extent of the induration at 48-96 hours with

a ruler to the nearest integer millimetre (e.g. 5mm, 12mm). Reactions over a certain threshold are deemed positive.

Enormous studies in the 1950s and 60s helped to ascertain the basic characteristics of the test. Distinct frequency distributions of reaction sizes were seen in 671,007 United States naval recruits with and without histories of TB exposure with the difference in the distributions approximating the distribution of reactions seen in 5,544 individuals with TB disease.(8) This suggested that positive reactions reflected previous TB exposure but that the test could not differentiate between individuals with past exposure and those with active disease.

The same studies(8) found that recruits from south eastern parts of the USA often had reactions to tuberculin regardless of TB contact; that these reactions were typically smaller than reactions seen in individuals with TB disease or a history of TB contact; and that the same individuals reacted to PPD-B. PPD-B was prepared from *Mycobacterium intracellulare*, a so-called environmental or non-tuberculous mycobacteria (NTM). The spatial differences in the distribution of reaction sizes was therefore interpreted as being a result of differences in exposure to NTM, which are found in, for example, soil or water.

'Non specific reactions' to the TST are also observed in children who have received a BCG vaccination. However, evidence suggests that, among children vaccinated in infancy, BCG explains very few false positive reactions of 15mm or more and that the BCG effect wanes with age, with little effect seen after 10 years of age (9,10).

Approaches to choosing TST cut points and ways of making inferences about the force of infection from TST survey data are discussed in Chapter 3.

### *Interferon gamma release assays*

In these tests, the patient's T cells are incubated *ex vivo* with antigens that are found in MTB but not found in BCG nor in most NTMs. Levels of interferon gamma are then measured in the supernatant or cells producing interferon gamma counted in an Enzyme-Linked ImmunoSpot (ELISPOT) assay.

IGRAs are a more specific test for MTB infection(11). They can also be performed in a single visit, which is a significant advantage. However, IGRAs are expensive and require phlebotomy. There is substantially less experience informing their use in epidemiologic studies. Another major problem is that the mode of the frequency distribution for positive reactions is very close to the cut point recommended by the manufacturer(12). This means that small differences in the choice of cut point can have a substantial impact on both inference and prevalence estimates.

### Tests for TB disease

Tests for disease include clinical diagnoses and symptom scores; radiological diagnoses, usually based on chest radiographs; and microbiological diagnoses. Obtaining a microbiological TB diagnosis is difficult in young children who tend not to expectorate sputum and who often have extra-pulmonary disease.

As this thesis focuses on infection rather than disease, I will not discuss these tests in detail. However, some are more sensitive (e.g. chest x-ray) and some more specific (e.g. reports of haemoptysis or a positive tuberculosis culture). The predictive

value of these tests will depend on the background prevalence of disease and also on patient characteristics.

### Molecular epidemiology

Another important toolset in understanding MTB epidemiology are strain-typing techniques, which can provide evidence to support or refute putative transmission events. Except in a few research centres, strain typing is generally not performed in high burden settings. These techniques require an MTB isolate so, typically, TB molecular epidemiology does not capture transmission to or from children. For the same reason, except in rare instances(13), molecular epidemiology does not capture TB infections that have not progressed to disease. TB molecular epidemiology is discussed in more detail in Chapter 2.

## The natural history of tuberculosis

The classic view of TB epidemiology is that a proportion of individuals exposed to a case of active pulmonary TB become infected with MTB, with those not previously infected converting from TST or IGRA negative and developing a positive test for infection. Thereafter, a proportion of those infected progress to active disease and a proportion remain latently infected though may 'reactivate' later. Much of the data supporting this sequence of events are from low incidence settings where individuals could be followed up after discrete episodes of exposure to infectious cases of TB. The literature is summarised well by Rieder(7). However, recent data suggest that this framework may not fully reflect MTB's natural history.

### Early clearance

Tests for MTB infection are known to have limited sensitivity(11) with the precise value calculated dependent on the cut point used and the choice of gold standard. Gold standard tests for TB infection include TB disease. Alternatively, one can assess tests' ability to detect a difference in infection prevalence between groups of individuals with more or less exposure to MTB. There is evidence that TST sensitivity might be further reduced in individuals with immune compromise resulting from, e.g., HIV or malnutrition, with limited evidence suggesting the same might be true with IGRAs(11,14,15).

Some have hypothesised that certain individuals might be able to clear MTB infection, probably via an innate immune response, without developing an adaptive immune response(16). Support for this suggestion comes from the discovery of genetic loci that predict TST reactivity(17–19) and interferon responses to antigens included in commercial IGRAs(20), plus the longstanding observation that some individuals with heavy MTB exposure remain TST negative(16). So-called 'early clearance' might be important epidemiologically. However, given it leaves no footprint, it is difficult to study.

### Conversions and reversions

Among individuals with positive tests for infection that do not progress to active TB, I am aware of no compelling evidence regards what proportion retain latent infection and what proportion manage to clear their MTB infection entirely. Demonstrating an absence of viable mycobacteria within an individual (living or deceased) would clearly be a challenge.

Instability of both TST and IGRA tests have been noted in individuals undergoing repeated testing(11,12,21) and the meaning of reversions debated.

Fine and colleagues(21) studied 64,225 tuberculin tests undertaken by the Karonga Prevention Study in a rural area in Northern Malawi. These data included paired tests undertaken on 6991 individuals at an interval of five years. They noted that the mode of the distributions of non-zero reactions increased with age from approximately 10mm in children to 15mm in adults. They also noted that TST reversions were common, particularly in BCG vaccinated under five year olds, and that the reversion rate settled to approximately 5 reversions per 100 person years by adolescence. A larger proportion of reactions in young children, particularly those who had recently received BCG vaccination, might be expected to be false positives. It was therefore suggested that the observed pattern in reversions could be, at least partially, explained by the initial positive tests being false positives.

Andrews and colleagues(12) studied adolescents in a community near Cape Town experiencing a very high force of infection. There were 5,357 participants of whom 3,236 underwent IGRA testing (QuantiFERON Gold) every three months. In the study, both TST and IGRA conversion were associated with an increased incidence of TB disease. A quarter of those who converted to IGRA positivity in the first year of follow up reverted to negative by the end of the second year. A quantitatively higher IGRA reading was less likely to revert, as might be expected. There was a suggestion that adolescents who IGRA reverted had the same TB incidence as those who IGRA converted and remained positive. However, these estimates were imprecise as there were very few cases of TB disease among individuals who IGRA reverted.

Both Fine and Andrews(12,21) note that test reversions may result in cross sectional studies underestimating the force of infection. This is discussed in Chapter 3.

### A continuum between infection and disease

Historically, MTB infection has been thought to consist of either 'latent infection' – not infectious and with little replication – or active disease – potentially infectious, if in the airway, and replicating.

This dichotomisation is probably a simplification of the biology. A number of groups have now proposed that MTB infection probably exists on a continuum (22–25). The proposed continuum goes from early clearance, through slowly dividing or 'latent' infection, then subclinical disease to symptomatic clinical disease. The further along the continuum, the greater the rate of MTB replication and the higher the bacillary load.

The evidence prompting this reassessment includes Positron Emission Tomography Computed Tomography (PET-CT) scans undertaken on individuals with active and latent TB. In very small studies, scans using radiolabelled glucose to detect metabolic activity showed occasional periods of activity in latent infection that were similar in appearance to metabolic activity seen in TB disease(26).

Regardless, it does seem likely that pulmonary infection on the more active end of this continuum would be required for infectiousness with peak infectiousness perhaps occurring just before the disease results in reductions in social contact, death or treatment initiation.

## Airborne transmission of *M. tuberculosis*

### Droplet nuclei and airborne transmission

MTB is one of a limited number of pathogens thought to be spread predominantly via the airborne route, although transmission via unpasteurised milk remains important in some communities(27) and transmission via direct inoculation onto broken skin or mucous membranes is possible(28). The airborne mode of transmission was predicted by Robert Koch(7,29) but only convincingly demonstrated in experiments at Johns Hopkins University more than seventy years later.

Prior to these experiments, William Frith Wells undertook a set of calculations predicting the behaviour of droplets expelled from the respiratory tract(30). He estimated that, whilst larger respiratory droplets would be expected to fall directly to the ground, smaller droplets (those 0.1 – 0.2 mm in diameter or smaller), would not. Their high surface area to volume ratio would result in water evaporating rapidly. Wells predicted that these smaller droplets would attain a sufficiently low mass before settling that they would remain suspended on air currents until either inhaled or ventilated out of the room. Wells predicted that infections transmitted via these small airborne particles, which he termed ‘droplet nuclei’, would behave differently to those spread via larger respiratory droplets.

> *'The epidemiological characteristics of droplet infection and droplet nuclei infection are by nature opposite. Droplet infection is essentially localized and concentrated while infection broadcast by droplet nuclei is dispersed and dilute.'*
>
> *William Frith Wells, 1934* (30)

To demonstrate that this route of transmission might be relevant in MTB transmission, Wells, Richard Riley, and others at Johns Hopkins University, set up a 'pilot ward' (31). In this ward, air from six rooms, each containing a tuberculosis patient, was exhausted through cages of rabbits or guinea pigs. The ventilation system was carefully calibrated with the aim of ensuring all animals were exposed to an equal volume of air from the ward. Incident infection was measured through vivisection and the counting of tubercles in the lung. The infection of guinea pigs and rabbits at some distance from source cases suggested airborne rather than droplet transmission.

This work did not rule out other modes of transmission, particularly droplet infection, also being important. However, three other pieces of evidence point towards airborne transmission being the dominant means by which MTB is transmitted between humans.

First, in separate experiments, Wells(32) and Lurie(33) demonstrated that aerosolised MTB in coarser droplets were less likely to result in tubercles in the lungs of exposed animals than MTB in finer droplets, with Wells observing no infections caused by droplets settling at more than one foot per minute. It is thought that larger

droplets are filtered in the upper airways and do not reach alveolar macrophages, the primary target in early MTB infection.

Second, descriptions of well-characterised outbreaks suggest exposure to shared air predicts MTB infection better than close or conversational contact. A widely cited example is the outbreak on the battleship USS Richard E. Byrd, where, in the mid 1960's, 350 navel personnel were exposed to a case of infectious TB for a period of at least six months(34). There were a number of secondary cases. Detailed data were available on social contact, sleeping arrangements and on the mechanical ventilation system. In the outbreak, the cabins in which the sailors slept and the extent to which these cabins were supplied by air from the cabins in which TB cases slept predicted tuberculin conversions well(34). Similarly, in a household contact study in the Gambia where, presumably, most household members had frequent close contact, MTB infection was strongly predicted by whether individuals slept in the same room or building as the index case(35).

Finally, 'cough box' experiments, recently undertaken in Uganda, suggest that MTB in droplet nuclei are the primary means of transmission(36,37). In these experiments, smear positive TB patients were asked to cough 'as frequently as was comfortable' for two five minute periods into a 'cough aerosol sampling system'(36). This system contained an Anderson six-stage cascade impactor that separated droplets by size before capturing them on culture dishes. Nearly all culturable MTB was found in droplets of less than 5 μm in diameter(36). These small droplets would be expected to remain suspended in the air rather than settling to the floor(30). Importantly, the number of these droplet nuclei containing viable MTB produced during sampling

predicted transmission to household contacts much better than smear status or time to culture positivity(37).

### Cough and other aerosol generating activities

Clearly, if droplet nuclei are the main means by which MTB is transmitted, aerosol generation is required for transmission to occur. The evidence for the role of cough in MTB transmission has recently been reviewed(38). Early studies showed that coughing was a very effective means of generating bioaerosols(39) and that there was substantial reduction in cough frequency after treatment initiation(40). More recently, an association has been described between patient reports of cough severity and transmission to household contacts(41). Some evidence suggests that speech and singing are effective aerosol generating activities(39,42,43) although the earlier studies(39,42) only measured aerosols arising in the oral cavity.

> *'The present writers would not wish their finding to be constructed as suggesting that singing should be proscribed. Not only would they prefer a world with tuberculosis to a world without song, but they recognise the relatively minor place occupied by this means of dissemination of airborne infection. Better that people gather to sing together than to shout at one another, even if the droplets produced are smaller and therefore occasionally more dangerous.'*
>
> *Loudon and Roberts, 1968*(42)

On going research suggests patients with TB disease may produce a higher concentration of droplet nuclei during normal tidal breathing than healthy

controls(44). This may be important, given most people spend more time breathing than coughing, speaking or singing.

### TB treatment and infectiousness

There is an active debate regarding how quickly TB treatment renders cases non infectious. The bulk of the evidence comes from animal studies and from studies looking at transmission to household contacts.

'Pilot wards' like that established at Johns Hopkins in the 1950s have been replicated in Lima, Peru, and in Witbank, South Africa. In the early studies, the source of transmission to animals was estimated from antibiograms and in the Peruvian studies spoligotyping was used. These are both low-resolution means of discriminating between MTB strains. There was a suggestion in all three sets of studies that the majority of transmissions were from patients who were yet to initiate treatment, patients who interrupted treatment, or patients who were on ineffective treatment as a result of drug resistance(45–49). It has been claimed that these studies show effective treatment rapidly reduces infectiousness(49). Broadly speaking, that is probably true. However, only a small number of patients were studied and the animals were in contact with air from the wards for several weeks. Had the animals been removed after a few days, it is possible that there would have been no infections in the guinea pigs exposed to untreated patients. In my view, these studies do not definitively establish a time on effective treatment after which infectiousness is abolished.

There are also limitations to the studies focussing on transmission to household contacts following initiation of treatment. There has only been one pertinent

randomised trial(50), which randomised smear positive adults in India to taking their first six months of treatment in hospital versus fully ambulatory treatment. The treatment in both arms was isoniazid and para-aminosalicylic acid. The trial compared outcomes among the household contacts of patients treated in hospital versus the household contacts of the ambulatory patients. It found no difference in TST at baseline, TST conversion over the subsequent six months or disease incidence at five years. However, it has been argued(51), given the high force of infection in the surrounding community, that a small excess of transmissions from the ambulatory patients might have been hard to detect. There have been several other studies comparing transmission to household contacts between patients who received ambulatory versus hospital treatment, all with reassuring results. However, these were non-randomised studies, meaning the patients selected for ambulatory treatment may have been less infectious than those treated in hospitals. Also, many of the household contacts were taking isoniazid preventive therapy. This literature has been reviewed by Menzies(51).

So, whilst it seems likely that a few days or weeks of effective treatment reduce infectiousness substantially, in my view current data do not clearly establish a time window after which patients can be considered non infectious.

## Risk factors for MTB infection

### The Wells-Riley Equation

The principal determinants of the risk of MTB infection in a defined space are described as a Poisson process by the Wells-Riley Equation(52).

$$\text{Risk of infection} = 1 - e^{-Ipqt/Q}$$

I is the number of individuals with infectious pulmonary TB

p is the rate at which the individual(s) at risk of TB breathe

q is the rate at which infectious individuals produce infectious 'quanta'

t is the period at risk

Q is the rate at which room air is either sterilised or replaced with uncontaminated air

Below, I discuss these variables in turn.

#### *TB prevalence*

Clearly, time spent breathing the same air as one or more individuals with prevalent infectious (i.e. pulmonary) TB is required for airborne MTB transmission to occur. The determinants of population level prevalence are TB incidence and disease duration. TB incidence is determined by the incidence of MTB infection and the proportion of individuals infected with MTB who progress to pulmonary disease. Disease duration is affected by case detection rates, the proportion of detected cases that receive effective treatment and mortality. At least in the pre-antiretroviral therapy era, HIV positive individuals with TB had a shorter duration of TB disease(53,54) resulting, presumably, from some combination of faster progression to illness (and thus

treatment), possibly faster diagnosis (as a result of a higher index of suspicion or regular contact with the healthcare system), and higher mortality. However, individuals with prevalent pulmonary TB who keep themselves to themselves do not present a transmission risk. Therefore, it is TB prevalence in putative sites of transmission that is the relevant parameter. For example, Rieder suggests that, as TB disease in several low burden settings is increasingly concentrated in older people, household MTB transmission (typically from parents to children) will become less frequent(7).

### *Respiratory rate*

Children breathe lower volumes of air per minute than women, and women, on average, breathe slightly lower volumes of air per minute than men(55). People engaging in physical activity breathe greater volumes of air than individuals at rest(55). Therefore, spaces in which individuals exert themselves (e.g. gyms or evangelical churches) might be riskier than e.g. libraries. Differences in respiratory volumes might – all else being equal – increase risk of MTB infection in adults, especially adult men, relative to children.

### *Quanta production rate*

Riley and colleagues described a quantum as 'the number of infectious airborne particles required to infect…[which] may be one or more airborne particles' (52).

In the animal studies, attempts were made to estimate q empirically. The Baltimore studies calculated a mean rate of quanta production of 0·62–0·82 and 1·25 quanta per hour in a group of hospitalised, HIV negative patients(46,56,57). The latter value is often assumed in models. In the Peruvian studies, an estimate of 8.2 quanta per

hour was obtained in a group of hospitalised, HIV positive patients(48). Higher estimates – e.g. 138 quanta per hour among those not wearing surgical masks – were obtained in the Witbank studies in a group of patients of mixed HIV status hospitalised for multidrug-resistant tuberculosis (MDR-TB)(58).

Much higher estimates of q have been obtained in people exposed to TB during aerosol generating procedures(59). However, estimates fitted to data on ventilation, social contact patterns and age structured data on tuberculin positivity in a South African community yielded an estimate of q of 0.89 per hour(60). Whilst this estimate was very sensitive to assumptions regards duration of infectiousness and the extent to which contacts were recurrent (i.e. with people with whom contact had previously been made), the relatively low value is interesting. A possible explanation is that undiagnosed individuals in the community may be less infectious than hospitalised patients, perhaps as a result of less advanced disease.

It is important to note that the Wells-Riley equation does not account for heterogeneity in infectiousness or susceptibility – both between individuals and within individuals over time. Heterogeneity in infectiousness is apparent not only in the animal studies described above (45,46,61) but also in population based studies(37,62–64).

Some of the drivers of heterogeneity in infectiousness are well described. It is well established that smear positive individuals are more infectious than smear negative individuals(65–67). However, recent household contact studies have yielded new insights. As mentioned above, only a minority of TB patients produce cough aerosols containing MTB and the quantity of these cough aerosols produced better predicts

infectiousness than smear status or time to culture positivity(37). Importantly, HIV related immune compromise in the index case has been shown to reduce transmission to household contacts. Contacts of index cases with a CD4 count of <250 cells/µl had 0.49 times the risk of TST positivity (95% CI 0.24–0.96) when compared with contacts of HIV negative individuals (little difference was seen between contacts of HIV positive people with higher CD4 counts and contacts of HIV negative people). This may be, in part, a result of shorter disease duration (see above). It may also be because HIV positivity is associated with smear negative TB and fewer cavities(68,69). These differences may be reduced with immune recovery in those taking antiretrovirals (ARVs) (68,69) although in one study(69) HIV positive miners with CD4 counts of <100 cells/µl had higher rates of smear positivity than those with higher CD4 counts, perhaps as a result of high bacillary load.

### *Duration of Exposure*

There have recently been attempts to measure duration of exposure to specific congregate settings in populations in Southern African communities(70–72). The length of time individuals are exposed to certain indoor spaces may or may not be modifiable. It is worth noting that, with long exposures, the cumulative risk of transmission may be high even in very well ventilated spaces(61). TB infection control strategies, such as the F-A-S-T strategy(73), have tended to focus on reducing the duration of infectiousness of individuals with TB rather than reducing the length of time over which susceptible individuals are exposed to relevant indoor spaces.

### *Ventilation*

In the Wells-Riley Equation, Q was defined as the rate at which the room was ventilated with 'germ-free air'. In the original paper, this was defined as a volume of clean air per unit time(52). However ventilation rates are often presented as the number of air changes per hour (ACH), with 1 ACH equivalent to approximately 63% of the volume of room air being replaced with clean air each hour.

A systematic review of observational and animal studies found 'strong and sufficient' evidence linking ventilation to TB risk(74) – a lot of the studies included in the review were descriptions of outbreaks. A recent household contact study in Uganda found weak evidence that TB disease in household contacts was less likely in better ventilated homes(75).

Alternative approaches to providing safe air can be described in similar language. For example, upper room germicidal irradiation (UVGI) is an effective means of killing aerosolised MTB(76) in which mycobacteria passing through the irradiated part of the room are killed by type C ultraviolet light. The impact UVGI on MTB transmission risk is often quantified by stating the equivalent number of ACH that would be needed to obtain the same effect.

I, t and Q are the three terms in the Wells-Riley Equation that are potentially modifiable. In spaces where the ventilation rate is much greater than the rate at which susceptible individuals breathe, transmission should be infrequent. It should be noted, however, that the Wells-Riley Equation assumes full air mixing. Even in well ventilated spaces, transmission may be possible in pockets of poorly mixed air or

prior to dispersal should individuals be directly exposed to a large number of infectious quanta.

### Exposure to public spaces

Two major motivations for undertaking the research described in this thesis were Suzanne Verver's molecular epidemiology study looking at intra-household transmission in Cape Town(77) and Olivia Horna Campos' studies from Lima, Peru(78–80). Molecular epidemiological studies estimating the proportion of TB disease that is a result of recent transmission within the home are the focus of chapter 2.

Olivia Horna Campos studied users of and individuals working in an informal public transport system(78–80). Some of the studies were small and limited adjustments were made for covariates. However, striking associations were observed between exposure to these congregate settings and TB or MTB infection. Among commuters with chronic coughs, individuals commuting to work using minibus taxis had 4.9 times the crude odds of pulmonary TB (95% CI 1.06–23.09) of those not commuting to work using minibus taxis(78). Among workers in the informal transport sector, those who had been in employment for more than two years had 15.7 times the age adjusted odds of a positive TST (95% CI 3.3 to 75.4) than workers who had been in post for less than two year(80).

There are other data linking exposure to public space with TB risk. I recently supervised a project in which we explored the association between working conditions and TB mortality using routinely collected data from 1890-2, 1900-2 and

1910-12 in England and Wales. We found that men working in crowded indoor environments had 1.43 times the rate of death from TB (95% CI 1.33-1.54) of men working outdoors or in less crowded indoor spaces, after adjusting for age, socioeconomic position and calendar period(81). However, for a number of reasons, including high mortality in men whose occupation we were unable to categorise, these results should be treated with caution.

In the South African gold mines, workers on the surface had 0.5 times the incidence rate ratio for TB disease of workers underground in an analysis adjusted for covariates including HIV and silicosis grade(82). Healthcare workers are known to have an elevated TB risk(83). This is likely, in part, to be a result of the high rates of undiagnosed TB disease found in healthcare facilities (84–87). However, part of the risk may simply be due to clinics and hospitals being congregate settings. The proportion of all MTB transmission that occurs in healthcare facilities is not known.

## TB epidemiology in KwaZulu-Natal and the wider region

The coastal province of KwaZulu-Natal (KZN) is South Africa's second most populace province, home to an estimated 10.3 million people according to the 2011 census(88). It is bordered by Swaziland, Mozambique and Mpumalanga province to the North, by the Eastern Cape to the South and, to the West, by Lesotho and Free State province.

The incidence of TB in KZN is exceptionally high with districts in the province reporting TB notification rates (all TB disease) of between 547 and 1097 per 100,000 per year (2013) (89). Capture-recapture studies suggest that notification data in KZN

are incomplete(90,91), so these data may underestimate the true burden. That KZN has the highest incidence of TB disease of any South African province(92) is probably explained by the province having a higher HIV prevalence than the other eight provinces. HIV prevalence in 2012 (all ages) was estimated to be 17.4% in KZN as compared to 12.6% in the country as a whole(93). This will equate to an adult HIV prevalence in KZN of approximately thirty percent. HIV prevalence is rising with the roll out of antiretroviral therapy (ART) (94). An estimated 2.3% of TB disease in the province is MDR-TB and, of this, 9.6% is extensively drug-resistant (XDR-TB)(95). These 2007 MDR-TB estimates are substantially lower than those reported by the District Health Barometer, which report the proportion of GeneXpert positive samples that have rifampicin resistance mutations(89). Patients, particularly those with complicated disease, may have more than one test for resistance and this will have inflated the estimates produced by the District Health Barometer.

Reliable recent data on the force of MTB infection in KZN are not available with the last tuberculin survey undertaken in 1974(96). Similarly, there has not been a recent TB prevalence survey in KZN (a national prevalence survey begins this year after some delays).

As outlined above, numerous factors may shape TB epidemics. These include HIV prevalence; social contact patterns, which are influenced by factors such as climate and household structure; the performance of the TB programme; the design of indoor spaces; and socioeconomic factors. Therefore, findings from one setting may not be reproducible elsewhere. The extent to which TB epidemiology in KZN is comparable to that in the rest of the region is an interesting question with implications for the generalizability of my results.

Other South African provinces – Eastern Cape, Northern Cape, Western Cape – have similar TB notification rates to KZN(92). However, the (all ages) HIV prevalence is much lower, particularly in Western Cape and Northern Cape (5.0% and 7.4%, respectively) (93). There is little published on TB epidemiology in Eastern Cape or in the sparsely populated Northern Cape. The force of infection in Western Cape has consistently been estimated to be higher than that measured in general community settings elsewhere in the world(2,12). The reasons for this are not clear though one might speculate that the cooler climate might play a role. Modelling work from Cape Town suggests that the spatial legacy of Apartheid era forced removals – leading to long commutes for the city's poor – are likely to adversely impact TB control(97). However, Gauteng province also has cold winters and informal settlements distant from sources of employment. Gauteng has a lower TB notification rate than the Western Cape despite higher HIV prevalence(92,93). I don't believe these observations have been fully explained.

In the Southern African region, there are a number of countries with similarly high estimated TB incidence to KZN. Lesotho, Swaziland, Namibia and Mozambique all had an estimated TB incidence of greater than 500 per 100,000 per year (2014) (98). TB disease in all four countries, as in KZN, is concentrated in HIV positive individuals. Botswana and Zambia have slightly lower TB incidence but, similarly, a large burden of HIV associated TB(98).

Swaziland is poorer than South Africa but, in terms of climate, culture and way of life, very similar to communities in the northern parts of KZN. In terms of climate and way of life, the same can probably be said for the southern parts of Mozambique. Lesotho

is mountainous, has cold winters and is more pastoral. I know less of the other countries.

Differences within countries may be more pronounced than those between countries. I think it might be reasonable to generalise my findings from a poor sub tropical rural community in KZN to rural communities with high TB incidence, and high HIV prevalence, in Swaziland, southern Mozambique, and perhaps elsewhere in Southern Africa. However, it would not be reasonable to use my data to make conclusions regards TB epidemiology in, for example, Durban or the wealthier parts of the Natal Midlands. Clearly, the proper test of generalizability would be to attempt to reproduce key findings in other communities.

## Conclusions

In this chapter, I have described evidence that airborne transmission is the dominant mode of transmission for MTB. I have described gaps in our knowledge of the natural history of the infection and the deficiencies of the tools we have at our disposal to study it. I have described risk factors for MTB infection with a focus on a growing evidence base that suggests exposure to public spaces is an important risk factor. I have reviewed the TB epidemics in KZN, South Africa and the wider region.

Important insights into MTB transmission have been obtained using the imperfect tools available to us. For example, Olivia Horna Campos' TST survey in public transport workers demonstrated clearly elevated risk associated with working in that environment(80). Similarly, a molecular epidemiology study from Los Angeles suggested that a substantial proportion of TB cases in that city (70% of cases where

the site of transmission was identified and 22% of all cases) were a result of transmission within three homeless shelters(99).

Similarly detailed insights into where transmission occurs in Southern Africa, based on empirical data, are urgently needed. Such insights would permit infection control and case finding interventions to be better targeted(3). In the chapters that follow, I attempt to fill that knowledge gap. My focus is on typical congregate settings rather than prisons and mines, which are known to be important sites of MTB transmission(83,100).

# 2. The Contribution of Intra-household transmission to Tuberculosis Disease, a Systematic Review and Meta-Analysis of Molecular Epidemiology Studies

## Background

In the mid 1980s, strain-typing techniques with better discriminatory power emerged. An early and, at the time, surprising observation was that the two individuals with active TB, resident in the same household, often had different strains of MTB(101,102).

Studies have subsequently attempted to strain type MTB isolates from all patients with active TB within defined communities and also to collect data on residence. Under a set of assumptions, these studies enable estimation of the proportion of all active TB (or all recently transmitted active TB) that is a result of recent MTB transmission between members of the same household. The first such study, conducted between 1993 and 1998 in Cape Town's northern suburbs, estimated that only 19% of active TB resulting from recent transmissions occurred as a result of MTB transmission between members of the same household(77).

Assumptions regarding household transmission of MTB may guide clinicians' treatment choices if no isolate is available from one of two linked cases. However, this thesis focuses on understanding MTB transmission in high burden settings rather than clinical management. Assumptions regarding household transmission of MTB also inform public health professionals' understanding of TB epidemiology and

therefore the design of programmes to interrupt MTB transmission. This is discussed in the final chapter of this thesis.

## Strain typing techniques

Historically, investigators attempted to differentiate between strains of MTB by comparing antibiograms or by using bacteriophages – these techniques have limited resolution.

### *Restriction Fragment Length Polymorphism (RFLP)*

The first higher resolution technique to emerge was restriction fragment length polymorphism (RFLP)(103). IS6110 is a 1,355 base pair sequence of DNA, which is found in most MTB genomes. Many strains harbour multiple copies. RFLP typing using IS6110 involves the following steps.

1. The DNA is 'digested' using a restriction enzyme – typically *PvuII*. This means that the enzyme severs the DNA chain wherever there is the specific DNA sequence it recognises. These sites can be found both within and outside IS6110 sequences.
2. Electrophoresis is used to separate the fragments by size.
3. A probe that binds to the IS6110 sequence is then used to visualise only the fragments that contain IS6110 sequences.
4. The pattern of bands is then compared with that from other isolates.

RFLP based typing of MTB was, for many years, the workhorse of TB molecular epidemiology. There are two major limitations of the approach. First, considerable laboratory capability and time are required to undertake RFLP based strain typing.

Second, some strains of MTB have few IS6110 copies. This reduces the discriminatory power of RFLP. Further limitations are considered below. It is considered good practice to use a second strain typing technique to add resolution when using RFLP to strain type isolates containing few copies of IS6110(104).

### *Multiple Interspersed Repetitive Units – Variable Number of Tandem Repeats (MIRU-VNTR)*

MIRU-VNTR was proposed as a less laborious method for strain typing MTB(105,106). The technique focuses on sections of the TB genome ('loci') that contain 'tandem repeats' – specific sequences of DNA that repeat themselves a variable number of times. Amplification of these sequences by polymerase chain reaction (PCR) and determination of the number of repeats based in the size of the amplicon allows a numeric classification of MTB strains with a sequence of numbers representing the number of repeats at each locus. The resolution of the technique depends on the number of loci used with, for example, 12 loci MIRU-VNTR having less resolution than 24 loci MIRU-VNTR(107).

### *Whole Genome Sequencing*

Rapidly reducing cost means that Whole Genome Sequencing (WGS) is likely to become the dominant strain typing technique for MTB. Typically, WGS does not sequence the whole genome as some sections are technically challenging.

Single nucleotide polymorphisms (SNPs) are the substitution of one nucleic acid for another – for example, substituting a guanine for a thymine – at a specific point in a DNA sequence. An understanding of the number of SNPs that separate distinct strains of MTB has been achieved through analyses of well characterised samples

(64,108–110). These have included samples from the same patient over time, samples from different anatomical sites within the same patient, samples from pairs of individuals with epidemiological links and samples from pairs of individuals without epidemiological links. These analyses have suggested that MTB evolves slowly, acquiring fewer than 0.5 SNPs per annum, although occasional bursts of more rapid evolution have been observed(111). WGS is able to discriminate between some isolates that are indistinguishable using 24 loci MIRU-VNTR or RFLP(64,108–110).

### Making epidemiological inferences from strain typing data

Epidemiological data and a number of assumptions are needed to make inferences about TB epidemiology from strain typing data. It is generally accepted that, if two individuals have different TB strains, it is unlikely that one infected the other. It is also assumed that, if two individuals have the same strain and a strong epidemiological link, it is likely that one transmitted to the other with transmission having occurred in the space in which those individuals spent most time together. The direction of transmission is then, sometimes, predicted based on which case came first or smear status.

Another common assumption in TB molecular epidemiology is that isolates of the same strain type obtained from a community within a few years of each other – so called 'clustered' isolates – represent ongoing MTB transmission. Unique strains are assumed to result from reactivation of strains that were circulating at the time the individual was infected. Clearly there are other potential explanations for unique strains including migration and low density sampling.

Potential biases that are particular to TB molecular epidemiology are complex(104,112,113) and, as they relate to this analysis, are addressed in the discussion.

### Meta-analysis of prevalence

Specific statistical issues(114) arise in the meta-analysis of prevalence, proportions and other numeric entities bounded by 0 and 1. Inverse variance weighting, which is commonly used, results in unsatisfactory results if some studies report prevalence close to 0 or 1.

First, the inverse variance method does not preclude upper or lower bounds for pooled confidence intervals lying outside the range 0 – 1. More importantly, estimated variance trends towards zero as prevalence approaches 0 or 1. As a result, studies of the same size can receive very different weightings.

Transforming prevalence to an approximately normally distributed variable unconstrained by 0 and 1 can address these problems. Meta-analysis of the transformed values is then undertaken using the inverse of the variance of the transformed proportions as the study weights. Pooled estimates and their confidence intervals can then be back transformed and reported.

Barendregt *et al* (114) argue that the best approach to the meta-analysis of prevalence is to use a double arcsine or Freeman-Tukey(115) transformation. The approach addresses the two statistical issues described above. They show that it performs better than a simple logit transformation which can result in too much weight being given to studies reporting proportions that are close to 0.5.(114)

## Objectives

1. To quantify discordance in strain type in pairs of individuals with active tuberculosis residing in the same household. This should be possible for all included studies.

2. To estimate the proportion of all MTB transmission in communities that occurs between members of the same households. This will only be possible for studies that attempted to strain type all MTB isolates in a defined community and which collected data on residence.

3. As a post hoc analysis, to use the method described by Crampin(116) to estimate the proportion of all TB disease in communities that is a result of recent transmission between members of the same household. This should be possible, with assumptions, for all included studies.

## Methods

### Search strategy and selection criteria

I searched for studies published after 1 January 1985 in MEDLINE, EMBASE, POPLINE, Global Health, Web of Science Conference Proceedings Citations Index (CPCI) Science and Web of Science CPCI Social Science and Humanities.

Detailed search terms are presented in Appendix 1. The strategy required reports to include standard TB terms AND terms encompassing household AND strain typing

terms. The strategy was developed using a set of test papers (64,77,116–118) and with advice from Caroline De Brun at the Royal Free Hospital Library, London, UK.

Titles and abstracts then full text papers were double screened for eligibility. In this, I had assistance from Dr Rob Aldridge of the Farr Institute, London, UK. Disagreements were resolved by consensus. The process was managed using EPPI Reviewer 4 (Institute of Education, London, UK).

The inclusion criteria were

- Empirical studies or systematic reviews focussing on TB in man.
- Published in English in or after 1985.
- Study describes ≥10 household pairs of co-prevalent active TB.
- Pairs strain typed using RFLP, MIRU-VNTR or sequencing.

When reviewing the full texts, we additionally excluded studies that only collected data on residence for isolates that were of the same strain. Whilst we allowed strain selection for the index case – for example, some studies only included drug resistant index cases – we excluded studies that employed strain selection for both index and secondary cases.

I required ten or more case pairs to be reported because I was concerned that small studies reporting strain concordance would be less likely to be published than small studies reporting strain discordance, given discordance, at least in the early days of TB molecular epidemiology, was considered more newsworthy.

I could not think of any reason why estimates of strain discordance from larger studies should be systematically different to that in the population as a whole. I thought the vast majority of TB molecular epidemiology studies would be published in English language journals. I could not think of any reason that studies excluded for being published in another language would be systematically more or less likely to report strain discordance than those published in English.

I hand searched the reference lists of included studies and contacted their corresponding authors to ask if they were aware of other studies. These additional studies were then reviewed by Rob Aldridge and me and included in the review if appropriate.

### Definitions

In most instances, defining 'household pairs' was straightforward. We accepted any sensible definition of household that was broadly consistent with the term as defined in Porta's Dictionary of Epidemiology – 'One or more persons who occupy a dwelling (i.e., a place that provides shelter, cooking, washing, and sleeping facilities); this may or may not be a family.'(119) We accepted studies that stated individuals were from the same household or lived together without explicitly defining what this meant. We excluded homeless individuals and individuals residing in institutions, such as homeless shelters, prisons or hospitals.

I decided not to set a maximum interval after which cases were no longer considered co-prevalent. Instead, I opted to stratify by the duration of the sampling interval (see below). The definition of co-prevalence used in each study is reported in Appendix 2.

The majority of reported cases were in households with two or fewer cases. However, some households contained more cases. I did not consider *a priori* how to deal with households containing more than two individuals with active TB. However, most studies had a sensible means of designating one case in such households the index case based on the timing of the diagnosis, smear positivity or pulmonary versus extra pulmonary disease (Appendix 2). We adopted the definition reported in each study and considered homes with *n* cases to contain *n-1* index-secondary case pairs, with the same index cases in every pair. Where this was not possible, we restricted analysis to households containing only two cases.

### Data extraction

I extracted the following variables.

- The country in which the study was conducted
- Primary strain typing technique (that with the highest resolution)
- Secondary strain typing technique(s)
- The period over which index case isolates were sampled
- The period over which secondary case isolates were sampled
- Whether the study was conducted in the context of a TB outbreak
- Whether there was pre-selection of index cases (e.g. drug resistant or smear positive disease)
- The number of discordant primary-index case pairs
- The total number of primary-index case pairs in which strain typing was available
- The number of discordant pairs (restricted to households containing two TB cases)

- The total number of pairs in which strain typing was available (restricted to households containing two TB cases)
- The proportion of all TB cases in the community exposed to a co-resident index case (if reported)

### Statistical analysis

Meta-analysis was undertaken, where appropriate, using the MetaXL plug-in (Epigear, Sunrise Beach, Australia) for Microsoft Excel (Microsoft Corporation, Redmond, USA).

I applied a Freeman-Tukey transformation to the data prior to meta-analysis and then back transformed the pooled effect estimates and their confidence intervals(114,115).

Strain concordance between pairs of co-resident individuals with active TB might be context specific. It might, for example, depend, upon household composition and construction, how well ventilated indoor public spaces are, social contact patterns and the performance of the TB programme. Whilst, ideally, the resulting heterogeneity would be dealt with by stratifying on such variables, in practice, these data are not available.

However, the extent to which context is important is unclear. Many molecular epidemiology studies are small and it is plausible that the probability that a small study is published is dependent on how 'interesting' the result seems. Discordance rather than concordance between household pairs of TB cases might be seen as more remarkable.

I present both fixed and random effects meta-analyses. Fixed effects meta-analysis should yield less biased results where the probability smaller studies are published depends on their findings. Random effects meta-analysis may better account for heterogeneity driven by differences between communities. Specifically, the standard errors and confidence intervals obtained from random effects meta-analysis may better reflect uncertainty resulting from sources of variation that cannot be dealt with by stratification(120). Differences in the results obtained by the two approaches may be informative.

The $I^2$ statistic(121) was used to quantify between study heterogeneity. I used a funnel plot to look for evidence of publication bias.

### Planned subgroup analyses

The following subgroup analyses were planned.

1. Stratification by TB burden.

There are two principal mechanisms by which strain discordance might come about (potential alternative explanations are reviewed in the discussion). There may be independent acquisition of recently transmitted MTB. Alternatively, one or more members of a household pair might reactivate their latent MTB infection. Exposure to TB in indoor public spaces will be less common in low burden settings. For this reason, co-resident individuals both being independently infected with different strains of MTB and progressing to disease might be rarer in low burden settings.

However, reactivation TB constitutes a greater proportion of TB disease in low burden settings and that may drive discordance in these communities(122,123).

To explore this, I stratified by WHO estimates(98) of national TB incidence at each study's midpoint. For example, if the first isolates were obtained in 1990 and the last isolates obtained in 1991, I took the mean of the 1990 and 1991 WHO incidence estimates.

In the absence of accepted international definitions of high and low TB incidence, I used the following strata: <20 per 100,000 per annum, 20-100 per 100,000 per annum and >100 per 100,000 per annum. Twenty per 100,000 is the European Centre for Disease Control threshold for a low burden country, based on the recommendations of a 2002 working group(124,125). One hundred per 100,000 per annum is an incidence threshold used by the World Health Organisation in their recent recommendations regarding LTBI treatment (6).

2. Stratification by the resolution of the strain typing technique.

The resolution of strain typing techniques is known to vary. Clearly, the resolution of 12 or 15 loci MIRU-VNTR is not as good as 24 loci MIRU-VNTR (123). RFLP and 24 loci MIRU-VNTR have been shown to have similar resolution (126,127), though modest concordance may suggest they evolve separately (126). WGS offers the best resolution and can discriminate between some isolates with the same RFLP or 24 loci MIRU-VNTR pattern(64,108–110).

Low resolution strain typing techniques could lead to strain discordance within households being underestimated. I will explore this by stratifying by the discriminatory power of the strain typing technique: low (12 and 15 loci MIRU-VNTR), moderate (24 loci MIRU-VNTR and RFLP) and high (WGS).

3. Stratification by duration of sample collection.

TB is a slow disease so molecular epidemiology studies of insufficient duration can result in incorrect inferences as linked cases are wrongly classified as unique (112,113,122,128).

In estimating strain discordance, one might expect duration of sampling to be important given progression to disease is most likely within the first two years after infection(7,129). The likelihood of a subsequent case being secondary to an index case would therefore be expected to fall as the interval between the two cases lengthened.

I will present an analysis stratified by sampling duration, using the categorisation from Houben and Glynn's systematic review (122): 0-12 months, 13-48 months and >48 months.

4. Stratification by pre-selection of TB cases with particular characteristics, such as smear positive or drug resistant cases.

Smear positive TB is known to be more transmissible (7). In many settings, there is delay in starting cases of drug resistant TB on adequate treatment, resulting in

greater potential for transmission (1). However, there are limited data that suggest drug resistance infers a fitness cost so drug resistant strains might therefore be less transmissible were effective treatment to be started promptly (130).

I will present an analysis stratified by whether there was any pre-selection of pairs according to the characteristics of the index case. Strain concordance among household pairs would be expected to be higher were the index case more infectious.

5. Restriction to studies not undertaken in outbreak conditions.

Outbreaks are, by definition, atypical and epidemiologists have noted the potential for 'outbreak bias' (131). For example, a high number of transmissions among members of a church congregation or within a large household might be deemed an outbreak. Clearly, there is the potential for bias in either direct here. I will, therefore, stratify by whether or not the study was undertaken in outbreak conditions.

### Planned sensitivity analysis

A small proportion of TB disease is in houses with more than two cases. In the primary analysis, I assigned one case in each pair to be the index case resulting in n-1 index-secondary case pairs. Given varying rates of disease progression and the possibility of transmission from smear negative individuals(132), this assignment may be unreliable. As a sensitivity analysis, I will repeat the main analysis only including households with two cases.

### Post hoc calculation of population attributable fractions

In the course of conducting this systematic review, it became apparent that I could apply the method described by Crampin *et al* (116) to estimate the proportion of all TB cases in the community attributable to recent transmission within households.

Crampin showed that the proportion of household index-secondary pairs showing strain concordance is equivalent to the risk difference percent. Therefore, the population attributable fraction approximates the product of the proportion of pairs that are concordant and the proportion of TB cases that were exposed to a putative source case within the home.

Note, Crampin required 'putative source cases' (index cases) to be smear positive and included non resident close family members (116). To enable the approach to be applied to all studies and to estimate household transmission only, I applied the definition of index cases used in each study (whether smear positive or smear negative) but did not include non-resident close family.

The proportion of cases in a community with household links to another TB case might conceivably be under or over ascertained. Under ascertainment might occur where the collection of epidemiological data is not thorough or follow up is of insufficient duration(133). Over ascertainment (relative to case finding in households containing only one case) might occur where household contact tracing programmes are strong or where diagnosis in one individual prompts other household members to test. I, therefore, planned to present population attributable fraction estimates using not only the estimate from the study (where available) but, if possible, also the

minimum and maximum estimates across all studies. Uncertainty in the proportion of pairs that were concordant was captured using binomial confidence intervals.

## Results

### Identifying eligible studies

An initial search was undertaken on 22 May 2013 and this was updated on 17 October 2014. In total, the searches identified 8806 references. Screening of titles and abstracts and then full texts yielded 15 papers(77,116,118,134–145) that met the inclusion criteria. A further 3 references(146–148) were obtained by other means. The process is summarised in a flow chart in Figure 1.

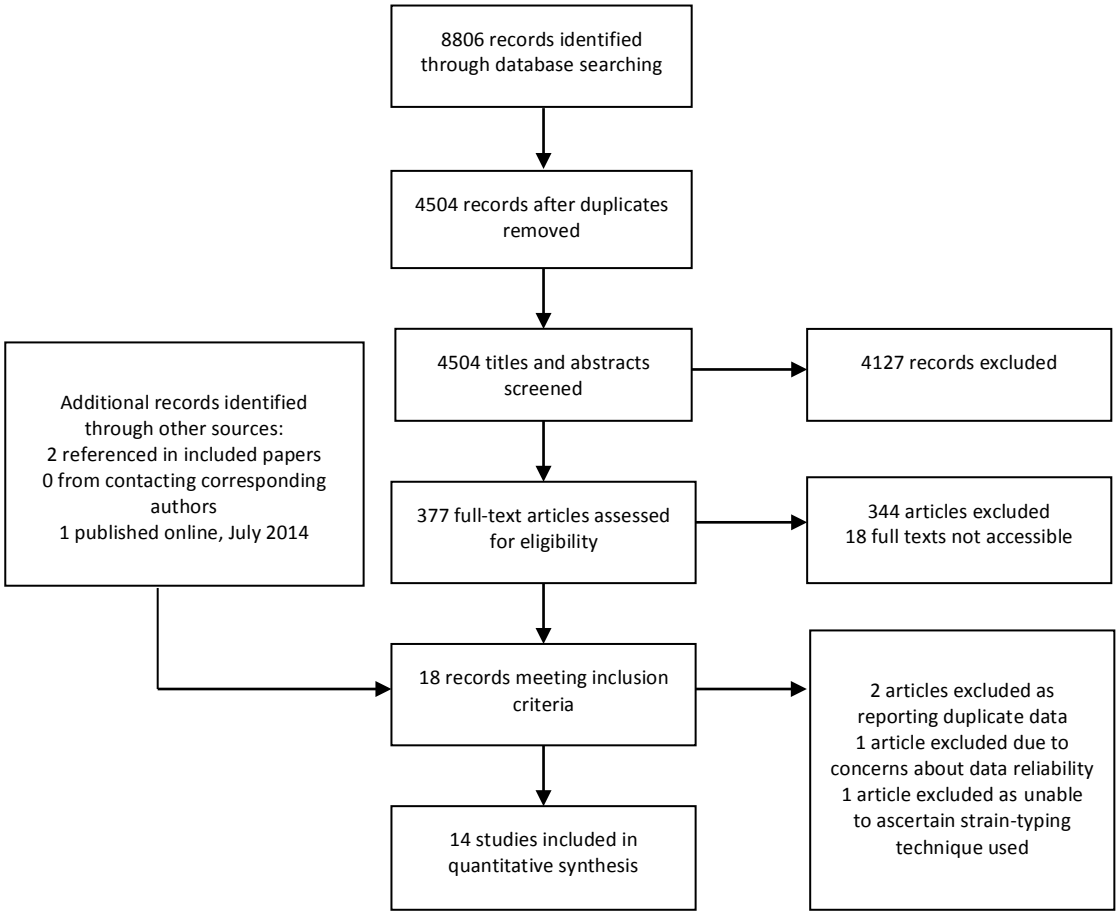
Three papers reported overlapping data from Cape Town(77,135,138). Only the paper(77) reporting data on the largest number of pairs was included in subsequent analysis.

One study was subsequently excluded due to concerns about data reliability (142). Conflicting data on strain discordance were presented in a figure and a table in the paper. Furthermore, ambiguity in the text made it unclear which cases were index cases. We were unable to contact the corresponding author.

A further study(136) was excluded due to ambiguity about the strain typing technique used. The paper stated 'isolates from index cases and their household contacts were compared using standard insertion sequence (IS) 6110 DNA restriction fragment length polymorphism (RFLP) and/or double repetitive element polymerase chain

reaction'. We were unable to contact the corresponding author to clarify what had been done.

*Figure 1. Flow diagram for the systematic review.*



### Data obtained from corresponding authors

We were directly provided with data by two corresponding authors.

Keren Middelkoop's study(148) of TB transmission in a township outside Cape Town included both the residents of formal 'plots' as well as people living in informal shacks. In the paper, it is stated that plots contained between 1 and 22 houses. Dr Middelkoop was unable to provide the data by household for the plots but was able to provide data on 13 co-prevalent pairs of TB cases residing in 13 shacks. The data on shack dwellers but not the data on residents of formal plots were included in the review.

We contacted Judith Glynn regarding a paper(116) describing RFLP typing of TB cases in Karonga District, a rural area in the North of Malawi. Prof Glynn offered to provide us with then unpublished WGS data on the same samples. These data have subsequently been published(149). Given the greater resolution of WGS, we opted to use these data in place of the RFLP data reported in the original paper. Whereas the paper groups household contacts and close family contacts, Prof Glynn was able to provide us with data on the household pairs only.

### Included studies

Our systematic review included data on 664 index-secondary case pairs, from 14 studies in 10 countries. The papers are described in summary in Table 1 and in detail in Appendix 2. The smallest studies(134,141) contained data on 11 pairs each and the largest contained data on 260 pairs(137).

Note that I intend to double extract these data prior to publishing this analysis in a peer reviewed journal. If this process identifies errors in data extraction, the results presented in the paper may differ somewhat from those presented in this thesis.

*Table 1. Summary of included studies.*

| Study | Country | Years | WHO mid study TB incidence (per 100,000pa) | Most sensitive typing method used | Duration of sampling (months) | Pre-selection of the index case | Outbreak conditions | Discordant / total index-secondary case pairs (%) | Discordant / total pairs in households with only two cases (%) | % cases exposed to putative source case in the home |
|---|---|---|---|---|---|---|---|---|---|---|
| Bennett et al, 2002(137) | USA | 1996-2000 | 8 | RFLP | 48 | No | No | 80 / 260 (31) | No data | Not presented |
| Behr et al, 1998(146) | USA | 1991-6 | 11 | RFLP | 72 | Pulmonary TB | No | 9 / 34 (26) | No data | Not presented |
| Martin et al, 2009(140) | Spain | 2002-6 | 20 | RFLP | 12 | No | No | 5 / 29 (17) | No data | Not presented |
| Borrell et al, 2009(141) | Spain | 2003-4 | 21 | RFLP | 24 | Pulmonary TB | No | 4 / 11 (36) | No data | Not presented |
| Augustynowicz-Kopec et al, 2012(143) | Poland | 2003-10 | 25 | 15 loci MIRU-VNTR | 93 | Pulmonary TB | No | 15 / 43 (35) | 13 / 34 (38) | Not presented |
| Inigo et al, 2003(139) | Spain | 1997-9 | 25 | RFLP | 36 | No | No | 0 / 14 (0) | 0 / 11 (0) | Not presented |
| Leung et al, 2013(144) | China (Hong Kong) | 1997-2011 | 95 | RFLP | 186 | MDR-TB | No | 9 / 19 (47) | No data | Not presented |
| Huh et al, 1995(134) | South Korea | Unclear | Likely >164 | RFLP | Unclear | No | No | 2 / 11 (18) | 2 / 10 (20) | Not presented |
| Buu et al, 2010 | Vietnam | 2003-6 | 179 | 12 loci MIRU-VNTR | 24 | Smear positive TB | No | 12 / 13 (92) | No data | Not presented |
| Verver et al, 2004 | South Africa | 1993-8 | 318 | RFLP | 72 | No | No | 57 / 94 (61) | 57 / 94 (61) | Definition of household not consistent with other studies |
| Sia et al, 2013 | Philippines | 2001-3 | 355 | RFLP | 36 | Smear positive TB | No | 10 / 16 (63) | 5 / 11 (45) | Not presented |
| Glynn, 2015 | Malawi | 1997-2010 | 393 | WGS | 168 | Pulmonary TB | No | 17 / 46 (37) | No data | 18.0 |
| Whalen et al, 2011 | Uganda | 1995-2004 | 440 | RFLP | 24 | Smear positive TB | No | 15 / 61 (25) | No data | Not presented |
| Middelkoop et al, 2015 | South Africa | 2001-10 | 948 | RFLP | 120 | No | No | 9 / 13 (69) [1] | 9 / 13 (69) | Definition of household not consistent with other studies |

1. These data are only for shack dwellers in this community.

### Meta-analysis of strain discordance

Meta-analysis without stratification revealed high levels of heterogeneity ($I^2$ 83%). The funnel plot (Figure 2), showed studies with outlying estimates of strain discordance (118,139) but did not suggest publication bias.

### *Stratification by TB incidence*

Figures 3 and 4 show fixed and random effects meta-analyses stratified by TB incidence at the each study's midpoint. Note, these are WHO national incidence estimates (as these were available for all studies) (98). Estimates local to the study site may differ (see Appendix 2). One study (134) did not state when samples had been obtained, only that the earliest patient started treatment in 1975 and the latest observed outcome was in 1992. The WHO incidence estimate for South Korea in 1990, the earliest year in the WHO dataset, was 164 per 100,000 per annum. Given the trend towards reducing incidence seen in subsequent years, I placed the study in the high incidence subgroup.

There was a modest trend towards greater discordance in higher incidence settings with, in fixed effects meta-analysis, pooled estimates of discordance of 0.30 (95% CI 0.25-0.36) and 0.27 (95% CI 0.19-0.35) in the low and moderate incidence studies versus 0.48 (95% CI 0.42-0.54) in the high incidence studies. The estimates from random effects meta-analysis were comparable but with wider confidence intervals reflecting the substantial residual heterogeneity within the medium incidence and high incidence strata. This heterogeneity means these pooled estimates, the overall pooled estimate, and any conclusions about the association between TB incidence and strain concordance should be treated with caution.

*Figure 2. Funnel plot showing the Freeman Tukey transformed proportion of household index-secondary case pairs with a discordant strain-type plotted against study size.*

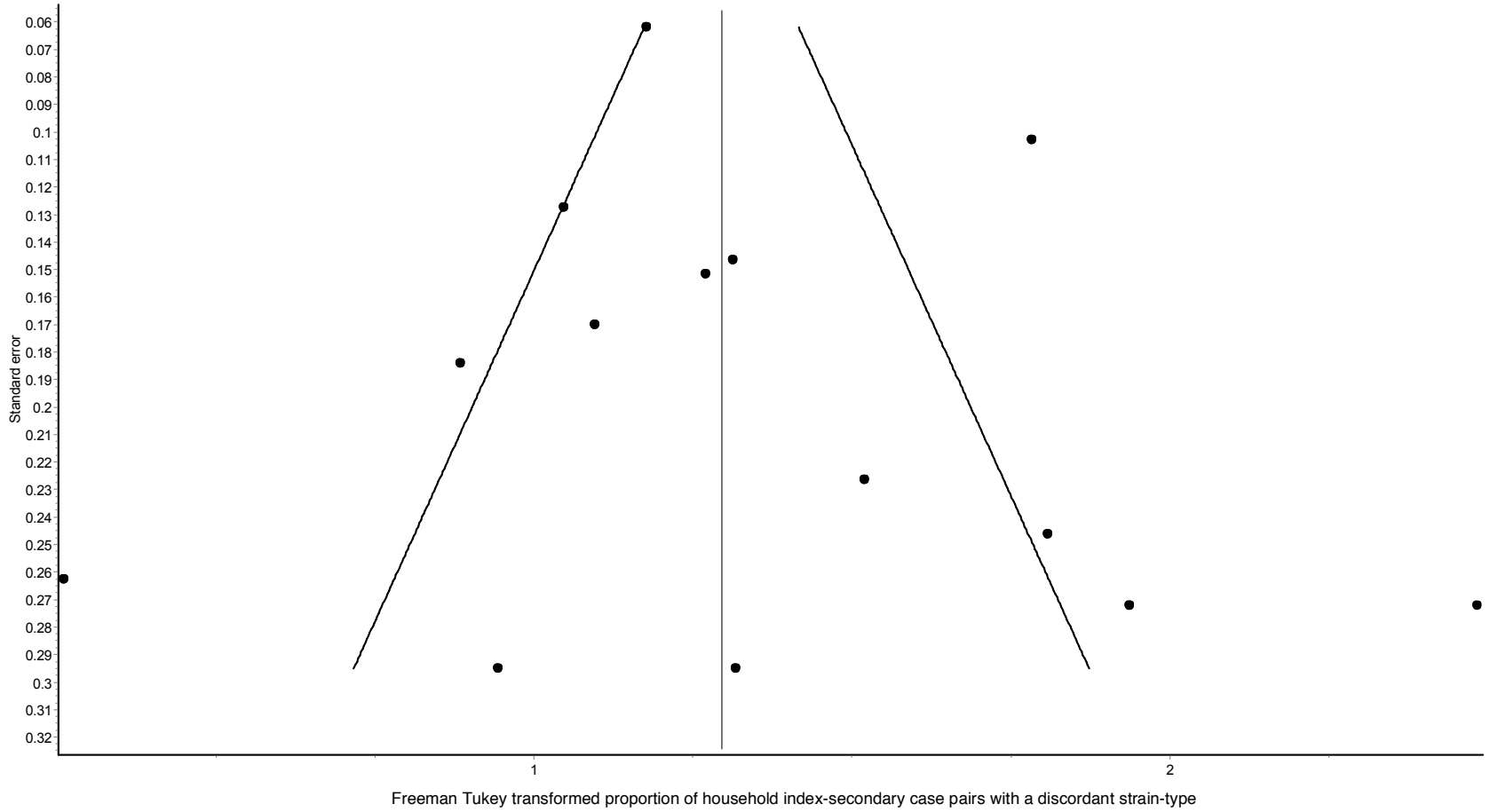


*Figure 3. Fixed effects meta-analysis, stratified by national mid study TB incidence.*

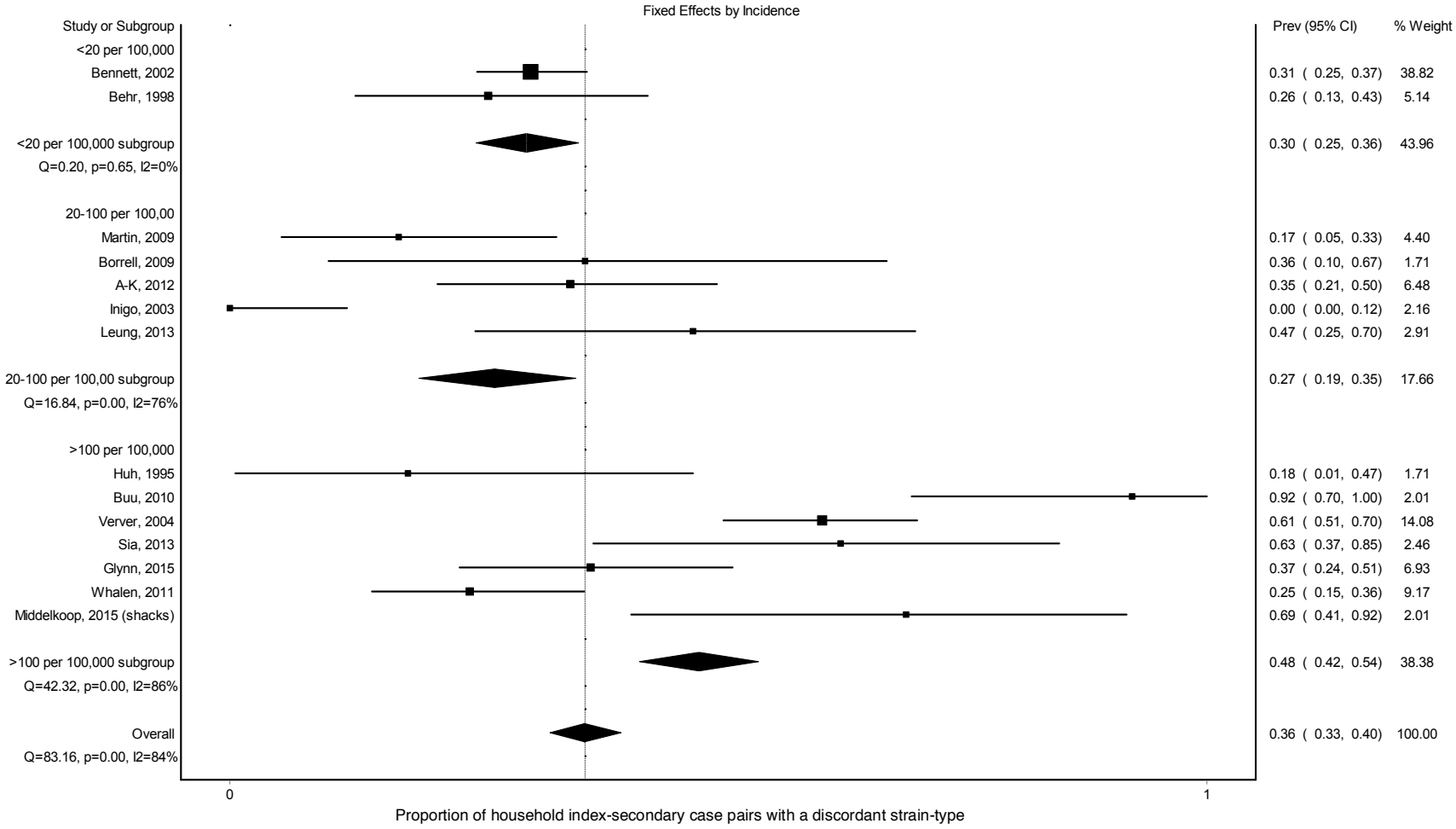
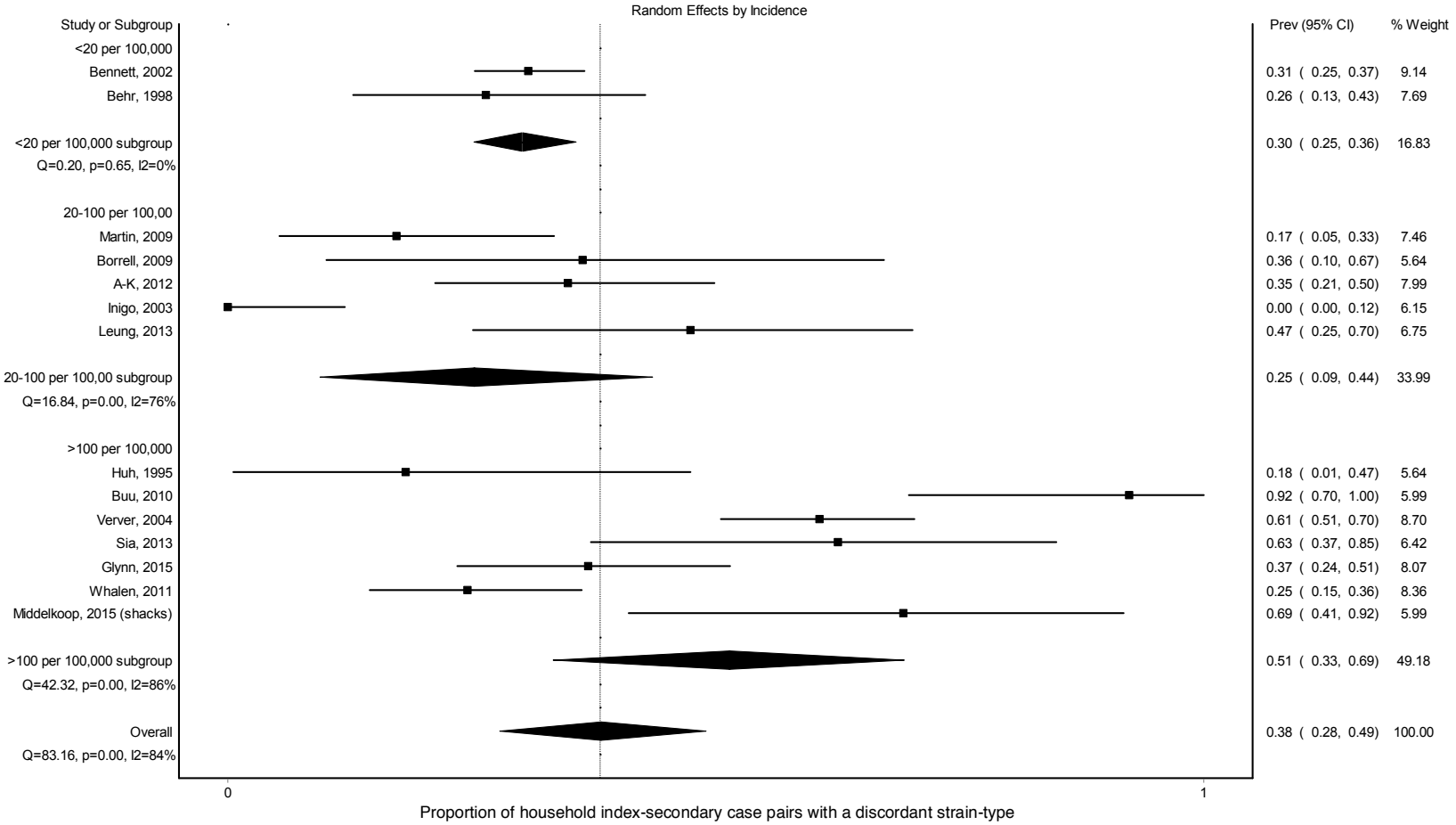


*Figure 4. Random effects meta-analysis, stratified by national mid study TB incidence.*



### *Stratification by the resolution of the strain typing technique*

Figures 5 and 6 show fixed and random effects meta-analysis stratified by the resolution of the strain typing technique. There was substantial residual heterogeneity in the low and medium resolution subgroups so the pooled estimates should be treated with caution. There was only one study reporting WGS data. The expected trend – less discordance with lower resolution techniques – was not seen. However, the vast majority of data was in the medium resolution subgroup (comprising RFLP and 24 loci MIRU-VNTR), with only 56 and 46 case pairs respectively in the low resolution and high resolution subgroups. Again, the confidence intervals were wider in the random effects meta-analysis. Particularly for the low resolution stratum, these wider confidence intervals seemed appropriate given the wildly discordant proportions reported in the Polish(143) and Vietnamese(118) studies.

*Figure 5. Fixed effects meta-analysis, stratified by the resolution of the strain typing technique.*

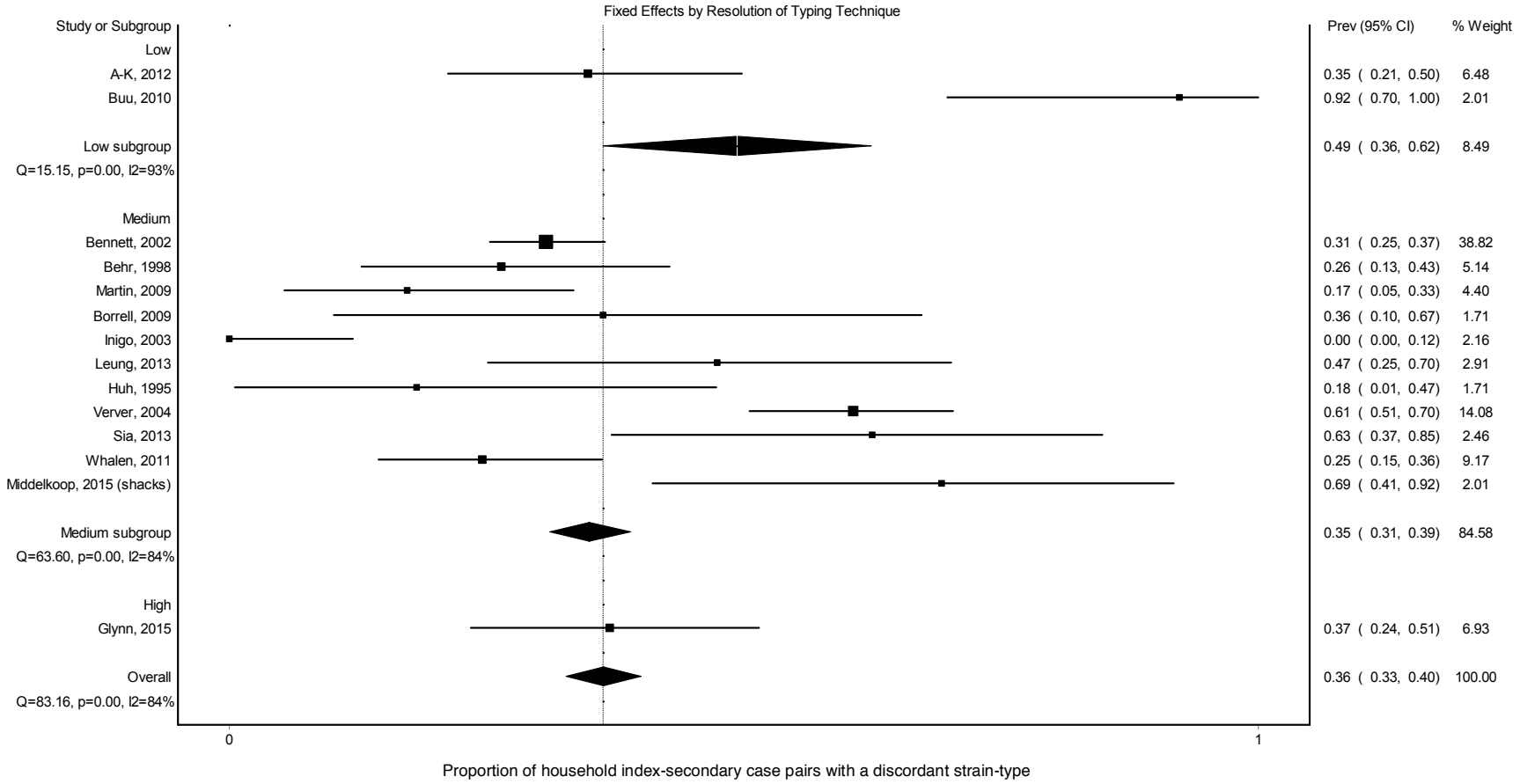
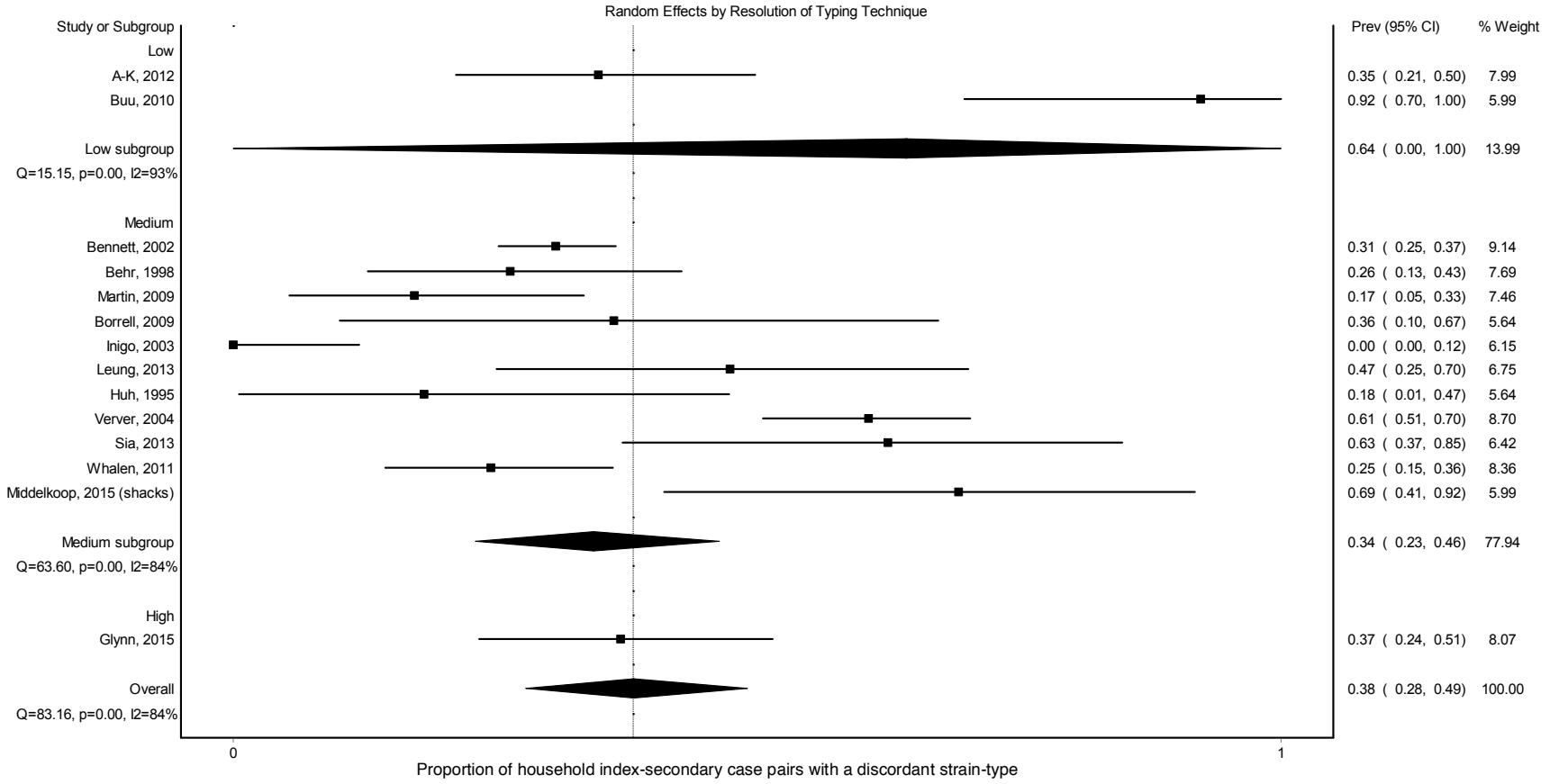


*Figure 6. Random effects meta-analysis, stratified by the resolution of the strain-typing technique.*



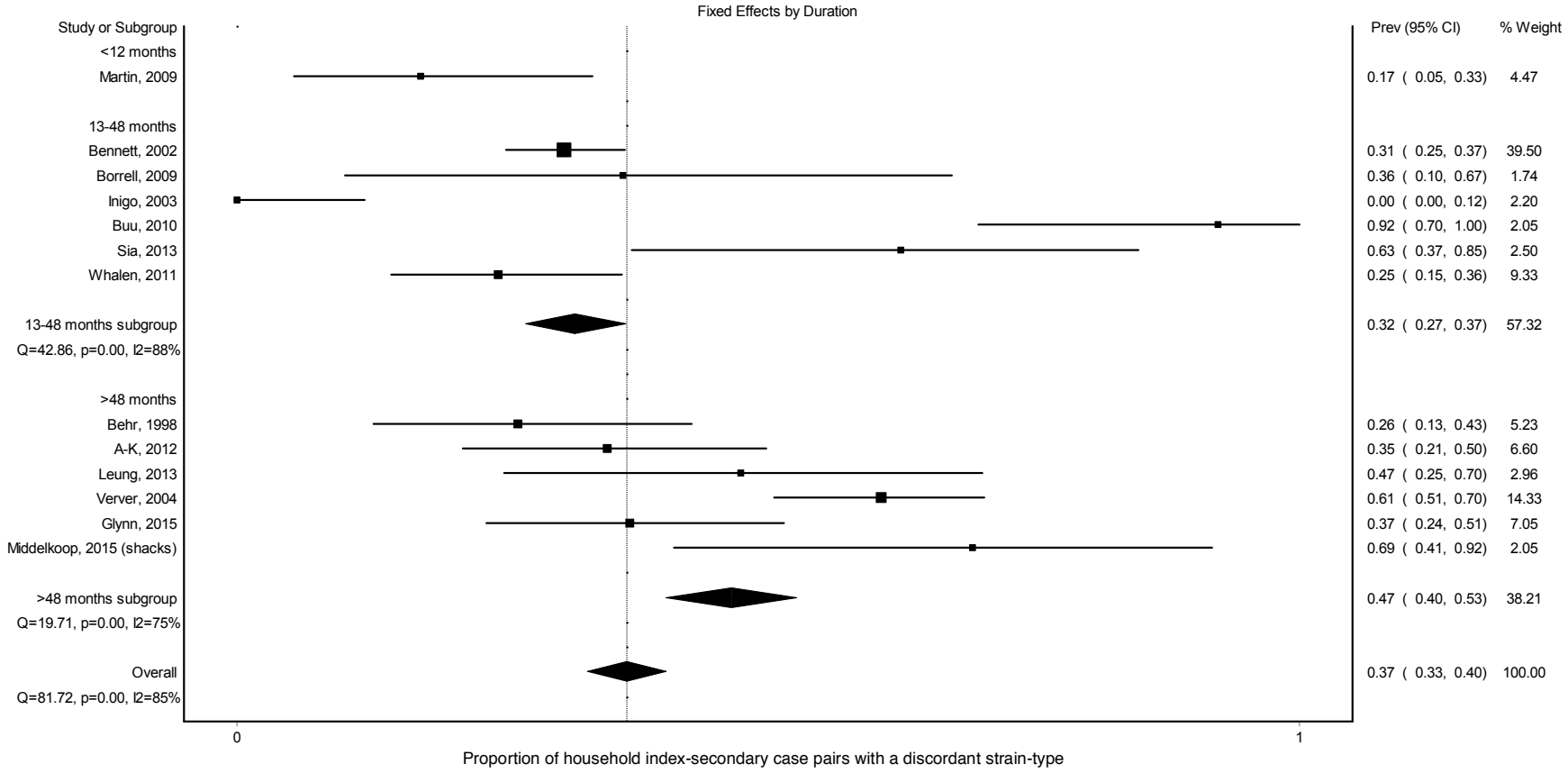

.

### *Stratification by the duration of sampling*

Figures 7 and 8 show fixed and random effects meta-analyses stratified by the duration of sampling. Note that we were unable to include one study in these meta-analyses (134) as the period over which sampling was conducted was not presented in the paper. The corresponding author did not reply to an email. Note also that some studies sampled for a longer period but only considered index-secondary case pairs if the isolates were obtained within a certain timeframe (118,140). For these studies we present the duration of the window, rather than the total duration of sampling. Details are presented in Appendix 2.

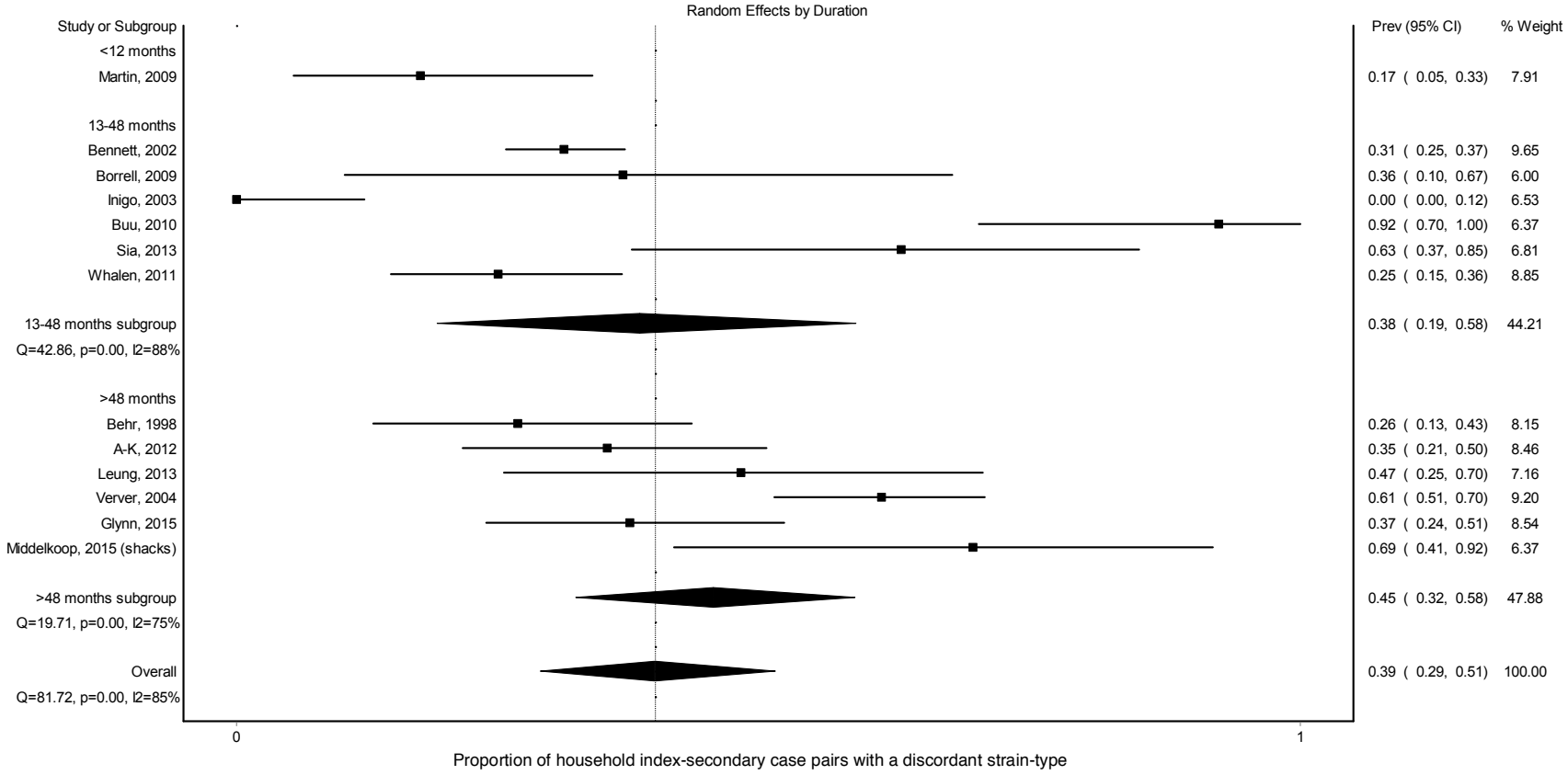
There were only 29 pairs in the <12 month stratum. There was a trend towards greater discordance with longer sampling. Pooled estimates of strain discordance from fixed effects meta-analysis were 0.32 (95% CI 0.27 – 0.37) and 0.47 (95% CI 0.40 – 0.53) in the 13-48 month and >48 month strata respectively. A similar pattern was seen in the random effects meta-analysis. This trend might be expected, given household transmission would be expected to end soon after diagnosis of the index case and given slightly over half of TB disease occurs within 2 years of infection (129). However, huge heterogeneity in the 13-48 month and >48 months strata should caution against making firm conclusions about trends from the pooled estimates.

*Figure 7. Fixed effects meta-analysis, stratified by the duration of sampling.*




.

*Figure 8. Random effects meta-analysis, stratified by the duration of sampling.*



### *Stratification by whether there was pre-selection of the index cases*

Figures 9 and 10 show fixed and random effects meta-analyses stratified by whether there was pre-selection of the index cases. To be included in this systematic review, isolates needed to have been obtained from both members of a case pair. It, therefore, seems likely that the majority of case pairs in papers not explicitly stating that they required index cases to have pulmonary TB involved at least one case of pulmonary TB. Again, huge heterogeneity should caution against making conclusions from the pooled estimates. However, it is somewhat surprising that discordance was not observed to be lower with a smear positive index case. Whilst smear positivity is an imperfect measure of infectiousness (36,37) there is a substantial literature demonstrating that smear positive disease is more infectious (65–67).

*Figure 9. Fixed effects meta-analysis, stratified by whether the index cases were pre-selected.*

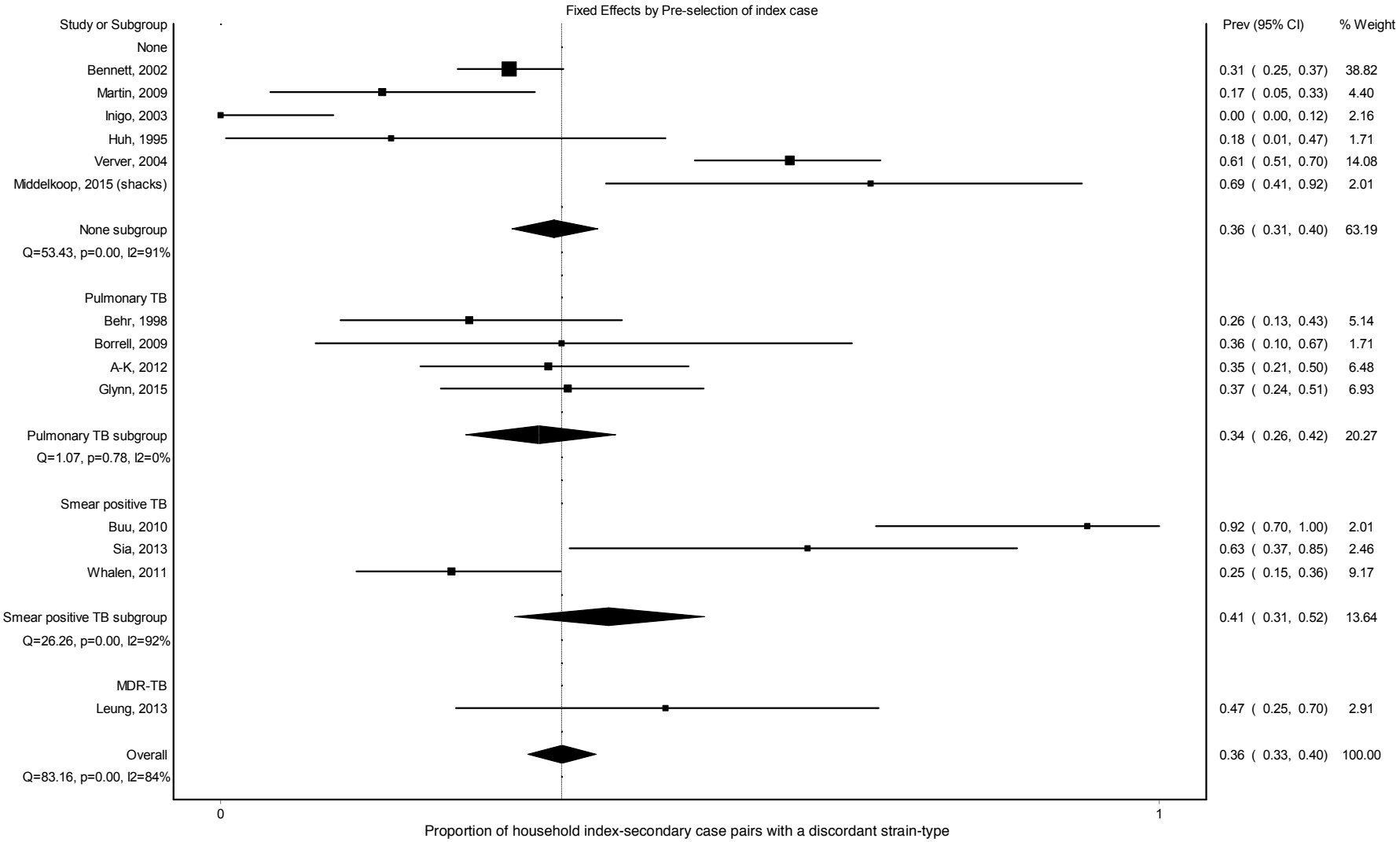
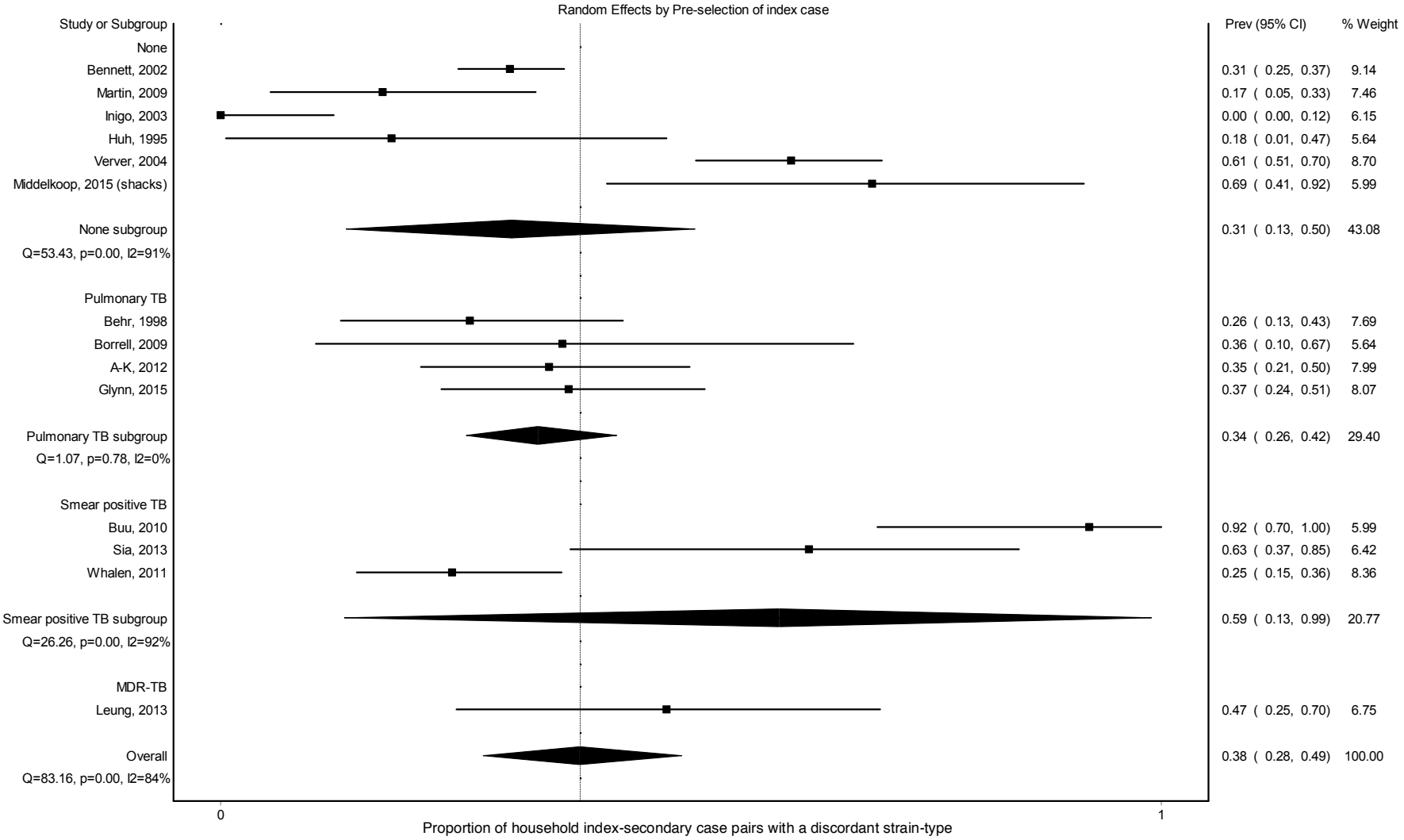


*Figure 10. Random effects meta-analysis, stratified by whether the index cases were pre-selected.*



### *Stratification by whether the study was conducted during an outbreak*

None of the studies were conducted in the context of a TB outbreak, so stratification on this variable was not necessary.

## Sensitivity analyses

The planned sensitivity analysis restricted to households only containing one pair of cases was not undertaken. This was because too few studies reported their data in a way that allowed this. However, it was notable that, except for one small study(145), discordance in the single pairs only analysis did not differ meaningfully from that reported in the main analysis (Table 1).

During data extraction, I realised that the definition of household adopted in one paper (77) was broader than that used in the other papers: 'We defined a household as a house and the associated informal dwelling or dwellings at the same address on the same plot of land'. Suzanne Verver confirmed in an email that their 'households' often contained more than one family living in different buildings on the same plot. She highlighted the potential complexity of the social relationships that might exist on a plot (150). A full set of meta-analyses excluding this paper is presented in Appendix 3. Exclusion of this, the second largest of the studies, resulted in a more modest trend towards greater strain discordance with increasing disease burden, at least in the fixed effects meta-analysis. It also resulted in a more modest trend towards greater discordance with a longer sampling duration.

### Estimates of the proportion of MTB transmission occurring within households

Two included studies presented estimates of the proportion of TB disease attributable to intra-household transmission (77,116) and two came close to doing so (118,148). These studies estimated the contribution of household transmission somewhat differently and one study (116) was later updated(149).

#### *Verver et al*

Verver presented data from Cape Town's northern suburbs (77). The study area is neither affluent nor is it an informal settlement. HIV prevalence is lower than in many South African communities (5.2% among women attending antenatal clinics).

Verver first calculated the number of 'clustered' cases within households, i.e. the number of cases sharing an RFLP type with another case in the same household. She then calculated the number of clustered cases within the entire community, i.e. the number of cases sharing an RFLP type with another case in the community. The number of secondary cases within households and within the community was then calculated as the number of clustered cases minus one – i.e. removing one index case from each community cluster and one index case from each household cluster. The proportion of cases resulting from recent transmission that were attributable to transmission within households was then calculated as the number of secondary cases within households divided by the number of secondary cases within the community.

Verver's estimate was that '19%...of transmissions in the community took place in the household.' She also presented a sensitivity analysis, in which the 30% of culture

confirmed cases for which no RFLP data were available were either assumed to all have unique fingerprints or all assumed to have clustered fingerprints. This resulted in estimates of the proportion of transmission occurring within households of 13 and 36% respectively.

It is worth re-emphasising the broad definition of household used in this paper (see above). Using a more typical definition of household, one would expect a lower estimate of the proportion of TB resulting from recent transmission attributable to transmission occurring within households.

### *Crampin et al*

Crampin presented data from Karonga District, a poor rural area in Northern Malawi (116). The area is relatively isolated, with little in and out migration (151). It has an HIV prevalence of approximately 13% (151).

Crampin used the approach described in the methods section of this chapter to calculate the proportion of all TB disease in the community attributable to 'recent transmission from identifiable close contacts'. In the 2006 paper, strain typing was by RFLP and the attributable fraction presented was the proportion of secondary cases attributable to 'recent transmission from identified smear-positive [putative source cases] in their families or households.' Non-resident family members were only included if first degree relatives or half siblings of the patient. The population attributable fraction was estimated to be 9% (or 13% if 86 reported but not confirmed putative source cases were included in the calculation).

Using the same approach and, mostly, the same cases, a paper by the same authors published since we undertook our search estimated attributable fractions using WGS strain typing data (149). They estimated that 8.1% of secondary cases were attributable to recent transmission from identified smear-positive source cases in their families. It is worth noting that, in this second analysis, the authors looked for transmission from smear negative putative source cases but found that the SNP differences between all smear negative putative source case – secondary case pairs were not consistent with transmission.

Note that, using these same WGS data, below I present estimates of the proportion of all TB disease attributable to transmission within the household, excluding transmission from non resident family members.

That Crampin and Glynn (116,149) calculated a lower proportion of TB attributable to household transmission than Verver (77) may be a result of methodological differences. As discussed, the definition of household was much broader in Verver. Furthermore, the denominator for Verver was clustered (i.e. recently transmitted) TB whereas the denominator in Crampin and Glynn was all TB.

#### *Middelkoop and Buu*

Both Middelkoop(148) and Buu(118) attempted to strain type all TB cases in a defined community and to collect data on their place of residence. They then made inferences about the extent of MTB transmission occurring between members of the same household.

Middelkoop presented data from another Cape Town community (148), an informal settlement with an adult HIV prevalence of 23-25% and a TB notification rate of approximately 2000 per 100,000 per annum. Middelkoop noted that whilst 45% of the adult patients with TB lived on a residential plot with one or more other adults with TB, only 8% of adults with TB lived on a residential plot with another case with the same RFLP type.

As noted above, residential plots in this community comprised between 1 and 22 houses. For the analysis of discordance (already presented) and the calculation of population attributable fractions (presented below), we used data supplied by Dr Middelkoop on residents of thirteen shacks and did not describe discordance among residents of the same plots. The number of adults living in a household with another case with the same RFLP type might be lower than the 8% who shared the same residential plot.

Buu presented data from a rural area of Vietnam, a country with an HIV prevalence of less than 0.5% (118). The household contacts of 1,442 smear positive index cases with MTB isolates were followed up passively over 24 months. Only 13 secondary cases were diagnosed. Of these, 12 had a strain type that was discordant from the index case (or 11 allowing for a single difference at one MIRU-VNTR locus). They, therefore, conclude 'most TB cases resulted from transmission outside the household'. Whilst passive follow up may have missed cases, this conclusion seems reasonable.

### Post hoc calculation of population attributable fractions

Population attributable fraction estimates are presented in Table 2.

Only one paper (149) presented data on the percentage of TB cases exposed to a putative source case in the household, so we were not able, as planned, to obtain lower and upper bounds for this proportion from across all studies. Given this paper aggregated household and community contact, we approached the corresponding author for estimates for household contact only. The proportion used here, 18%, is the proportion of TB cases with a putative source case with whom they were 'living in the same household while the person was ill'.

As argued above, there is likely to be relationship between duration of follow up and strain discordance with secondary cases occurring long after the index case less likely to be a result of intrahousehold transmission. Therefore, multiplying the proportion of cases linked to another case in the household from a study with a long duration of follow up (Glynn(149) sampled over 168 months) by the proportion of pairs with strain concordance from a study with a short duration of follow up may lead to an overestimate of the contribution of intrahousehold transmission.

*Table 2. Post hoc estimates of the proportion of all TB resulting from recent within household transmission.*

| Study | Country | Midpoint incidence (per 100,000 per annum) | Discordance (95% binomial CI) | Percent of secondary cases exposed to a putative source case in the household | Population attributable fraction, given reported exposure (%) [1] | Population attributable fraction, applying the exposure estimate from Glynn (%) [2] |
|---|---|---|---|---|---|---|
| Bennett(137) | USA | 8 | 30.8 (25.2 – 36.8) | No data | N/A | 12.5 (11.4 – 13.5) |
| Behr(146) | USA | 11 | 26.5 (12.9 – 44.3) | No data | N/A | 13.2 (10.0 – 15.7) |
| Martin(140) | Spain | 20 | 17.2 (5.8 – 35.8) | No data | N/A | 14.9 (11.6 – 16.9) |
| Borrell(141) | Spain | 21 | 36.4 (10.9 – 69.2) | No data | N/A | 11.5 (5.5 – 16.0) |
| A-K(143) | Poland | 25 | 34.9 (21.0 – 50.9) | No data | N/A | 11.7 (8.8 – 14.2) |
| Inigo(140) | Spain | 25 | 0.0 (0.0 – 23.2) | No data | N/A | 18.0 (13.8 – 18.0) |
| Leung(144) | China, Hong Kong | 95 | 47.4 (24.4 – 71.1) | No data | N/A | 9.5 (5.2 – 13.6) |
| Huh(134) | South Korea | 164 | 18.2 (2.3 – 51.8) | No data | N/A | 14.7 (8.7 – 17.6) |
| Buu(118) | Vietnam | 179 | 92.3 (64.0 – 99.8) | No data | N/A | 1.4 (0.0 – 6.5) |
| Verver(77) [3] | South Africa | 318 | 60.6 (50.0 – 70.6) | No data | N/A | 7.1 (5.3 – 9.0) |
| Sia(145) | Philippines | 355 | 62.5 (35.4 – 84.8) | No data | N/A | 6.8 (2.7 – 11.6) |
| Glynn(149) | Malawi | 393 | 37.0 (23.3 – 52.5) | 18.0 | 11.3 (8.6 – 13.8) | 11.3 (8.6 – 13.8) |
| Whalen(147) | Uganda | 440 | 24.6 (14.5 – 37.3) | No data | N/A | 13.6 (11.3 – 15.4) |
| Middelkoop(148) | South Africa | 948 | 69.2 (38.6 – 90.9) | No data | N/A | 5.5 (1.6 – 11.1) |

1. Concordance multiplied by the percentage of TB cases exposed to a putative source case.
2. Concordance multiplied by 18.0%, the proportion of TB cases exposed to a putative source case reported in Glynn et al(149).
3. Note, the definition of household in this study was broader than that in other studies.

## Discussion

### Main findings

Whilst there was huge heterogeneity between studies, strain discordance among household pairs was observed to be common in several both high and low burden communities. There was a modest trend towards greater strain discordance between household pairs in higher burden settings and with a longer duration of sampling. However, substantial residual heterogeneity should caution against drawing firm conclusions about trends from these pooled estimates.

Four studies(77,118,148,149) attempted to collect data on residence and strain typing data from all cases within defined communities. These studies were undertaken in different though all relatively high burden settings. The parameters estimated varied between the studies. However, these studies suggested that, in these high burden settings, recent MTB transmission within the home makes a modest contribution to TB disease with one study suggesting that transmission in the home explains a small proportion of TB disease resulting from recent infection(77).

Finally, for all of the studies, we calculated *post hoc* the population attributable fractions for the proportion of all TB disease in the community that results from recent infection within the household. Only one study(149) reported the proportion of TB cases exposed to an index case in the home. To obtain attributable fraction estimates for the other populations, we multiplied that proportion (18%) by the estimate of strain concordance from each study. Given the huge differences in study design and in the characteristics of these communities, applying one estimate from rural Northern Malawi to thirteen very different populations may not be valid.

Specifically, given the long duration of follow up in the Malawian study, multiplying that estimate by estimates of concordance from studies of shorter duration is likely to result in an overestimate of the proportion of TB that results from recent transmission in the household. However, even assuming no strain discordance, as was observed in one small study(139), this approach resulted in a maximum estimate of the proportion of all TB disease resulting from recent MTB transmission in the home of 18%. It is notable that in the study from Vietnam(118), household secondary cases were only observed to arise from 0.8% of index cases over 24 months of follow up. Whilst the passive case find approach used in that study may have missed cases and some additional cases might be expected to arise after 24 months of follow up(129,133), the Vietnam data do suggest that exposure to a putative source case in the household may be much less frequent in some communities.

## Interpretation

### *Strain discordance*

Studies from a wide variety of settings showed that finding co-prevalent cases of TB in the same household but harbouring a different strain is not unusual and, in some settings, may be as common as strain concordance. Strain discordance may be driven by different things in high versus low burden settings.

In high burden settings, and in pockets of disadvantage in low burden settings, high rates of ongoing transmission outside the household may explain strain discordance. Data suggest that, on average, the yield of household contact tracing varies relatively little between high and low burden settings (133). However, clearly the risk of transmission in the community will be very different given differences in prevalence of

almost two orders of magnitude. Independent acquisition of MTB in the community should therefore be relatively much more common in high burden settings.

In low burden settings, TB as a result of reactivation of latent MTB infection comprises a bigger proportion of disease than in high burden settings (122). If one or more of a household pair had TB as a result of reactivation, strain discordance would be expected. However, TB disease in low burden settings is a rare event. Therefore, two cases of TB of different strain types within the same household should occur very infrequently. That discordant household pairs are noted with considerable frequency even in low burden settings is therefore curious. Possible explanations include the existence of pockets of high transmission within low burden communities – there are certainly data that suggest such pockets exist (99,152). Another potential reason for discordant household pairs would be household clustering of risk factors for TB infection or reactivation. These might include HIV infection, deprivation, overseas birth or travel, or time spent in spaces, such as prisons or homeless shelters, in which there are higher rates of MTB transmission than in the general population.

### *Within household transmission*

Only four studies had attempted to collect data on residence and strain type isolates from all TB cases in a defined community. All were from high burden settings, all estimated slightly different parameters, and all found that recent transmission in the home made a modest contribution to either TB as a result of recent transmission(77) or all TB disease(118,148,149). This is consistent with a growing body of research.

TB prevalence surveys find most cases of disease occur singly in households. For example, in a household survey in rural India in the early 1960s, Narain(153) found

‘Over 80% of the total number of infected persons, in any age-group, occurred in households without cases. Moreover, persons suffering from the disease were mostly found singly in households. Thus infection and disease were scattered throughout the community.’ He speculated that this might be due to ‘the comparatively early diagnosis of cases in a household survey.’ Clearly, prevalence surveys might find individuals earlier in their illness. However, even in routine contact investigations, only a small minority of MTB patients are found to have another case in the home. In a recent systematic review(133), the yield of household contact investigation was estimated at 3.1% (95% CI 2.0-4.4) in high income countries and 3.1% (95% CI 2.1-4.5%) in low and middle income countries.

Given most TB cases have not had recent exposure to another TB case within the home, it should perhaps come as little surprise that the proportion of TB disease resulting from recent transmission within the home is modest.

Molecular epidemiology can only study TB infection that progresses to disease. However, there are empirical and modelling studies that suggest that much TB infection occurs between rather than within households.

Data from case control studies suggest that the association between household TB contact and TB infection in children depend on the TB burden in the community(154) and on the age of the child(154,155), with the association between household exposure and TB infection less pronounced in older children and in settings with a greater TB burden. This probably mostly reflects greater exposure to TB outside the home(70,156,157) rather than reduced transmission from infectious household contacts.

There are data suggesting that, in some communities, poor household construction might be associated with a reduced risk of household transmission of TB, perhaps reflecting better ventilation in these less well-constructed buildings.(75,158,159) However, the modest effect sizes seen for the association between the quality of household construction and TB suggest that, even if – as seems likely – poor household construction is commoner in communities with high TB burdens, this could not explain the attenuation of the association between household contact and TB infection observed with rising TB burden.

Finally, there is mathematical modelling(60) from the same high burden community near Cape Town studied in one paper (148) included in this systematic review. The model suggested that 84% of MTB transmission occurs outside one's own household with school (for children), workplace (for adults) and public transport being key sites of MTB transmission(60). The modelling used Rudnick and Milton's adaptation of the Wells-Riley Equation(160), social contact pattern data(70) and carbon dioxide measurements to estimate risk of infection in each setting. The approach requires strong assumptions about full air mixing, the prevalence of infectious individuals in each space and the absence of heterogeneity in both infectiousness and susceptibility. The findings from this model regards household transmission do, however, seem consistent with the studies using alternative methodologies that I have described.

### *The proportion of all TB resulting from within household transmission*

Many of the observations I made previously are relevant here. It is worth, however, noting two additional things about this *post hoc* analysis.

First, the low proportions of all TB resulting from recent transmission in the home can partly be explained by reactivation TB appearing in the denominator. This is particularly true for the studies from low burden settings(122) where the majority of TB will be reactivation disease.

Whilst 83% of TB disease has been estimated to occur within five years of MTB infection(129), MTB has a variable incubation period and disease resulting from household contact can occur decades later(161). Clearly, what I have termed reactivation TB may be a result of remote intrahousehold transmission. Whether or not disease resulting from remote rather than recent transmission is more or less likely to be a result of exposure in the home will depend on changes in TB epidemiology in the intervening years. It is, however, also possible that infectious dose predicts primary progression versus infection and subsequent reactivation. Given household exposure would be expected to result in a high infectious dose, the contribution of household transmission to disease resulting from recent transmission may differ from that resulting from more distant transmission.

Second, in studying Table 2, it is worth noting the extent to which this result is driven by the low proportion of secondary cases exposed to a putative source case. Substantial caution is needed here, as this proportion is taken from a single study from rural Malawi. However, for the reasons outlined above, one might expect this proportion to be lower in, at least some, other settings. Therefore, regardless of biases and assumptions needed in analysing the molecular epidemiology data, it seems likely that, in some but perhaps not all settings, recent transmission in the home makes a modest contribution to the overall burden of TB disease.

### Limitations

There are a number of important limitations to this analysis.

Despite the comprehensive approach taken, we were only able to find data on 644 case pairs. Whilst data were reported from ten countries, many studies were small with a majority of pairs reported from only three countries: the United States (260 case pairs, one study(137)), South Africa (107 case pairs, two studies(77,148)) and Uganda (61 case pairs, one study(147)). The Vietnamese study(118) was also large and the limited number of household case pairs detected in that study, of itself, informative.

Nearly all included studies used RFLP or 24 loci MIRU-VNTR, making it difficult to make inferences regards the impact of the resolution of the strain typing technique on our estimates. In areas with limited diversity among circulating strains, or with substantial in-migration from such an area, co-prevalent TB cases sharing a common isolate may not reflect transmission. We know TB evolves slowly(64,108–110) and that WGS can distinguish between some strains that appear concordant with RFLP or 24 loci MIRU-VNTR (64,108).

It is also possible that differences in one RFLP band or at one MIRU-VNTR loci might be consistent with transmission. To evaluate whether this was likely, it would have been informative to see data on the extent of differences between discordant strains within households as compared to discordant strains within communities. Few studies presented such data.

In these analyses, I assume that most individuals with TB harbour a single strain. Mixed infections do occur and in certain, largely high burden, settings may be found in 10-20% of individuals with TB(162). In individuals with disease as a result of household MTB transmission but also a second strain, the strain transmitted in the home may not be the one isolated. Where this occurred, concordance and the extent of household transmission might be underestimated. However, it is worth noting that such individuals also harbour an isolate that wasn't transmitted in the home and there is – by the same mechanism – the possibility that transmission outside the household could also be underestimated. Only if the MTB transmitted in the home was systematically more or less likely than the other strain to be the one isolated would inference be biased across the study population – this seems unlikely.

There is an assumption in these analyses that two individuals with co-prevalent disease and the same TB strain, residing in the same household, must result from transmission within the home. However, clearly people living together may also attend the same public spaces and could have both been infected by a third individual, e.g. at church. These analyses might, therefore, slightly overestimate the proportion of TB disease resulting from intrahousehold transmission.

It should also be noted that by 'household', here we refer specifically to one's own household. In communities where people spend more time in each other's homes, there may be between household transmission that occurs in people's homes. For these analyses, such transmission would be considered between household or community transmission.

There are potential biases that apply to Verver(77) but not to the other studies, given Verver used as her denominator clustered TB cases. Inference from TB molecular epidemiology is unusual in that the sampling fraction can effect the parameters of interest directly rather than via selection bias(104,112,113). This is because missing cases can result in clustered cases being wrongly classified as unique. Low sampling density, migration and cases occurring before or after the study period can all result in relevant links being missed.

Verver thought that misattribution of clustered cases as unique would be more likely to happen in households given household clusters tended to be smaller than community clusters. In her study(77), most household clusters consisted of only 2 cases. I would agree with Verver's assessment but note that contact tracing or disease in one member of a household prompting others to seek healthcare might reduce the impact of this bias.

Finally, there are limitations that are inherent to all molecular epidemiological analyses. Even in good laboratories, a proportion of 'positive' MTB cultures will be a result of laboratory contamination or clerical error with one review estimating a median false positive rate of 3.1% (163). Whilst such errors could bias results in either direction, one might expect a false positive culture resulting from laboratory cross contamination to be less likely than a truly positive culture to be concordant with another sample from the same household. However, large scale laboratory contamination events can result in large numbers of false positive cultures, all with the same strain type. Such an event could result in an overestimate of strain concordance within household pairs – e.g. if both 'cases' were actually diagnosed as a result of false positive MTB cultures.

### Generalisability

Many studies did not include children (Appendix 2) and, among those that did, children will have been underrepresented given the difficulty in obtaining isolates for typing from younger children. Given differences between adults and children in social contact patterns, infectiousness and susceptibility to TB, these analyses should not be used to draw strong conclusions about the role of household transmission in childhood TB.

It should be noted, however, that the basic epidemiology described above suggests, at least in older children, that household transmission might play a similarly limited role in transmission to children, particularly in high burden settings. It should also be noted that Schaaf(138) reported on adult-child pairs in Cape Town's northern suburbs (most if not all of these individuals will have been included in Verver(77)). The 35 children with RFLP type available reported by Schaaf had a median age of 4 years (range 0.5 to 15 years). Only 19 of the 35 children with typing data lived in a household with another bacteriologically confirmed TB case. In 12 of the households, the adult(s) had isolates with the same RFLP type as the child; in 5 of the households, the strain type was discordant; and in 2 households, there was no adult isolate available for strain typing. An analysis stratified by the age of the child was not presented. Whilst caution should be exercised in extrapolating from so few cases, these data suggest that, at least in some high burden settings, transmission outside the home may also be an important determinant of childhood TB.

Few studies defined what they meant by a household. Household composition and the construction of dwellings probably varied considerably across the communities

studied, which might explain some of the residual heterogeneity. Caution should be exercised in generalising these findings to communities that are very different to those contributing substantial amounts of data to this review – America(137), urban South Africa(77,148), urban Uganda(147), Vietnam(118) and rural Malawi(116,149). We specifically excluded studies of the homeless population. Our results therefore cannot be used to make inferences about MTB transmission in hostels and other more communal residences.

### Pooled estimates

Where there is substantial heterogeneity, there is a choice. Systematic reviewers can either elect not to present pooled estimates or pooled estimates can be presented with the necessary caveats. I elected to do the latter as I felt that the trends across strata were of interest. I trusted that readers would treat pooled estimates from strata with substantial residual heterogeneity with the necessary caution. I acknowledge the risk that these pooled estimates might be misused and understand the case for not presenting them.

It might be argued that random effects models better reflect the uncertainty where there is substantial heterogeneity, i.e. sources of variation that cannot be accounted for through stratification. However, Greenland and O'Rouke (120) note that random effects models replace 'a doubtful homogeneity assumption…with a fictitious random distribution of effects'. In reality, neither set of assumptions is ever likely to be true. Methodological work is ongoing to develop valid approaches to obtaining pooled estimates in the context of substantial heterogeneity(164). Models that enabled systematic reviewers to present pooled estimates from heterogeneous studies with

confidence intervals properly reflecting the additional uncertainty resulting from the heterogeneity may, in the future, provide a better solution to this problem.

## Conclusions

There is a growing body of basic epidemiology and modelling that suggests MTB transmission outside the household is an important determinant of TB infection, particularly in high burden settings. There are substantial limitations to the data presented here and more data from a wider variety of settings are needed before these findings can be confidently generalised. However, the results of this systematic review and meta-analysis are consistent with the basic epidemiology and modelling in suggesting that MTB transmission outside the household is an important determinant of TB disease. It is known that healthcare facilities(83,86,165), prisons(100), mines(82,166) and homeless shelters(99,152) are important sites of MTB transmission. A fuller understanding of the contributions that transmission in these and other public spaces make to TB epidemiology is needed. The implications of these findings for TB control are discussed in the final chapter of this thesis.

# 3. *M. tuberculosis* infection in the Africa Centre Demographic Surveillance Area – a descriptive analysis

## Background

### The Setting

The Africa Centre for Population Health has been operating a demographic and health household surveillance programme in KwaZulu-Natal, South Africa, since 2000. From 2003, adults in the programme have been offered an HIV test annually. Buildings in the surveillance area are geolocated, allowing spatial analyses to be undertaken. Details of the platform and the data collected are available in published cohort profiles(4,167).

The surveillance area is 438km$^2$ in size, bounded to the east by the N2 national road and to the west by Hluhluwe-Imfolozi Game Reserve. The area does not include Mtubatuba, a small market town on the other side of the N2. It does include KwaMsane, a township on the N2, and a number of smaller settlements.

The population of the surveillance area numbers approximately 90,000 people around two thirds of whom are resident at any one time. There are approximately 11,000 homesteads and the population density varies substantially from 20 residents per km$^2$ in more rural areas to 3000 per km$^2$ in KwaMsane(4). Across the region, rural homesteads tend to be dispersed rather than concentrated in villages.

The surveillance area is in Umkhanyakude District, which was the second most deprived district in South Africa according to the 2011 Census(168). The principle

sources of income are state pensions and waged employment(4). In 2013, 72.7% of households drank piped water or water from a borehole and 81.1% of households were on the electrical grid (Africa Centre household surveillance data).

Healthcare is provided by seven nurse run primary healthcare clinics (PHC), six within the surveillance area and one in Mtubatuba. The clinic in KwaMsane is larger than the others with a small laboratory on site. Most people access healthcare at their nearest clinic(169) with substantial reductions in healthcare usage as distance to the nearest clinic increases.(169,170) Nurses working in PHC initiate most TB and HIV treatment. The nearest district hospital is in Hlabisa, which is 24km west of the surveillance area, on the other side of Hluhluwe-Imfolozi Game Reserve.

### Disease Burden

Between 2004 and 2011, the proportion of all HIV positive people receiving antiretroviral therapy (ART) increased from 0% to 31%(94). This increase in ART coverage was associated with an 11 year increase in life expectancy(171) and consequently an increase in adult HIV prevalence, from 21% to 29%(94).

The TB notification rate in Umkhanyakude District was 878 per 100,000 per annum in 2013 (168) and 730 per 100,000 per annum in 2014(89). In the communities living close to the Africa Centre, 76% of TB diagnoses are in HIV positive people.(172) In 2007, 4.8% of notified TB cases in Umkhanyakude District were multi-drug resistant TB (MDR-TB).(95)

To my knowledge, the last TB prevalence and tuberculin surveys in KwaZulu-Natal were conducted in 1974(96). The paper reported an 0.8% prevalence of culture

positive TB in adults. The team estimated an annual risk of MTB infection of 1.4% in children aged 1-17 years, assuming a third of reactions of 5-9mm and all 10mm or above represented MTB infection. At this time, the provision of health services for people who were not white was poor and BCG coverage was low. Only 28.7% of those under 18 years in the study had BCG scars despite a catch-up campaign providing BCG vaccination children at school enrolment. At the time of the survey, HIV had not arrived in Southern Africa. A comparison between the data from 1974 and my results will be made in the discussion.

Verbal autopsies undertaken by the Africa Centre(173) suggest that pulmonary TB is the biggest cause of premature adult deaths in the surveillance area (causing an estimated 48.8% of deaths in those aged 15-49 years). 'HIV/AIDS' (which will include extra-pulmonary TB) is the next most common cause of adult deaths(173). However, caution is needed in interpreting these verbal autopsy data as TB often presents with non-specific symptoms.

## Interpreting TST survey data

There are a number of challenges in making inferences about the incidence of MTB infections from TST survey data. There are choices to be made with regards to the cut point used. Also, assumptions are required to make inferences about the force of infection, i.e. the annual risk of MTB infection (ARTI), from data on the prevalence of TST positivity in children.

### *Choosing TST cut points*

Interpreting TST data is challenging because BCG coverage, NTM exposure, TB burden and the distributions of TST reactions attributable to each differ between

populations. Overlap between the distributions of reactions means that, in the presence of non-specific reactions, picking cut points is challenging(174).

Clinicians might choose a more sensitive cut point (e.g. 5mm). However, often researchers wish to find the cut point that best discriminates between non-specific reactions and reactions that are a result of MTB infection – the position of this cut point will vary between populations. Inevitably, there will be some misclassification. However, with a discriminatory cut point, prevalence estimates will be broadly correct. The number of TB reactions misclassified as non-specific reactions will be approximately the same as the number of non-specific reactions misclassified as TB(174).

A number of standard thresholds appear in the literature – e.g. ≥10mm (commonly used) (175) and ≥15mm (176). Based on data from Tanzania, a threshold of ≥14mm has also been used, multiplying the prevalence obtained by 1.22 to correct for the proportion of cases that were missed when using this threshold in Tanzanian populations with little NTM exposure(177).

An alternative approach(178,179) assumes that the distribution of true positive reactions is approximately normally distributed and that reactions at and above the mode are a result of MTB infection. The number of individuals with MTB infection is then calculated as the number at the mode plus twice the number with reactions bigger than the mode. This is the ‘mirror method’. There is also a ‘fixed mirror method’, fixing the ‘mode’ at 17mm, regardless of the true distribution in the data, based on a mode observed in patients with TB disease(180). Prevalence obtained by

mirror methods is very sensitive to the mode chosen and, in some TST data, a clear mode is not apparent.

Statistical methods have been developed to determine, in a given distribution of TST reaction sizes, the most likely underlying distributions of non-specific reactions and reactions resulting from MTB infection. The approach uses mixture analysis to test which underlying distributions best fit the data observed. Typically, two component models are fitted to the observed distribution of non-zero reactions, with one component representing MTB reactions and one representing non-specific reactions. Constraints can be applied – for example, fixing the number of components. Also, initial values for the mean and variance of component distributions can be set to approximate those seen in similar settings.

Simpler approaches to mixture analysis involve fitting two normal distributions to the data(181,182). This can result in unrealistic predictions if the tail of the NTM distribution suggests a large number of reactions of less than 0mm. More sophisticated approaches(183) test whether normal, log normal or Weibull distributions best fit the data, avoiding this problem. Different distributions can be fitted to each component.

The mixture analysis approach is appealing in that it enables an analysis plan to be pre-specified where little is known about the frequency of MTB and non-specific reactions within the population of interest. However, the approach may perform poorly where there is considerable overlap between the distributions of the MTB and non-specific reactions (common in settings with an intermediate prevalence of MTB infection, particularly if the mode of the MTB reactions is low)(183). It also performs

poorly where the sample size is small, and where measurements of induration size have been inaccurate(183). An evaluation of mixture analysis using data from the Karonga Prevention Study found the number of components that gave the best fit to the data varied with age and that some predicted distributions were hard to explain – for example, an NTM component that was only seen in individuals under the age of 30 years(184).

Given real uncertainty in many populations regards the optimal cut point and the limitations of mixture analysis outlined above, some authors(185,186) have opted to present sensitivity analyses using each of the approaches highlighted above. This is the approach I have taken in this thesis.

In the spatial analyses in this chapter (discussed below) and in the risk factor analyses presented in the next chapter, the choice of cut point is – in many ways – a choice between statistical power and specificity. Researchers looking for associations between putative risk factors and MTB infection might opt for a very specific cut point (e.g. 15mm), sacrificing power for accuracy. However, often researchers do not have sufficient data to do this whilst retaining sufficient precision. Standard approaches to determining cut points that best balance power (sensitivity) and specificity (e.g. the use of receiver operating characteristic curves) are not possible (I believe) if the underlying distributions of reaction sizes for a particular population are not known. Where there are two sources of data on the same individuals (e.g. TST and IGRA results) that both predict MTB infection, there is interest in using model based approaches, such as latent class analysis, to estimate both the probability of MTB infection in individuals and the sensitivity and specificity of specific TST and IGRA cut points(187). However, usually two sources of data are

not available. The solution to these problems is, in my view, more likely to come from improvements in diagnostic tests than from statistical innovations. Statens Serum Institut are currently trialling a tuberculin that contains antigens found in MTB nut not in NTMs or in BCG. Early data suggest that the test has good specificity and sensitivity comparable to IGRA, though performs poorly in individuals with advanced immunocompromise(188,189). Poor performance in advanced immunocompromise should not greatly affect inference in school based studies, particularly following the successful roll out of prevention of mother-to-child transmission programmes.

### *Inferring incidence from prevalence*

Tuberculin school surveys have long been a cornerstone of TB epidemiology. Given MTB infections in young children must have occurred between birth and the test date, they can be used to make inferences about recent transmission within communities. Prevalence can be converted into an annual risk of TB infection (ARTI) using the approach proposed by Nyboe(190).

$$ARTI = 1 - (1 - prevalence)^{1/mean\ age}$$

This approach assumes the force of infection has not changed over the lifetime of those tested, which is a strong assumption(174,190). If the force of infection is falling or rising, assuming the force of infection does not differ with age (another strong assumption), the estimate of ARTI will best reflect the force of infection approximately midway between the children being born and being tested. Age assortative mixing (70,71,157) and the fact that children do not frequent some community settings mean that caution should be exercised in using risks calculated in children to make

inferences about transmission to adults. Furthermore, from a single survey, it is not possible to unpick age from cohort effects(191).

For these reasons, plus the problems with the interpretation of conversions and reversions (detailed in Chapter 1), calculations of ARTI obtained using TST data should be interpreted as a proxy rather than a precise measure of the force of infection in a particular population.

### Spatial Scanning Statistics

Most infectious diseases display a level of spatial heterogeneity or dependency, meaning that the proximity of a case makes the outcome more (or, on rare occasions, less) likely. This spatial dependency may result from person to person spread or from spatial clustering of risk factors.

One challenge in looking for clustering of cases is that, in a given area, there are an infinite number of potential centroids and cluster sizes possible. A common means of dealing with this problem is by using a permutation test, such as the Kulldorff spatial scanning statistic(192). This approach scans the area under consideration so that all potential cluster centroids and sizes are evaluated. The single cluster assigned the highest likelihood of being a true cluster (the test can look for areas of high density only or for areas of both high and low density) is evaluated first. Significance testing is then undertaken by generating a large number of replicate datasets. In each new dataset, there are the same number of cases and controls and the same spatial locations. However, the cases and controls are randomly assigned to alternative locations. The significance of the cluster can then be ascertained from the proportion of replicate datasets in which an equivalent or more extreme degree of clustering is

observed at the same location. The process can then be repeated, evaluating the next most likely (non overlapping) cluster.

### Spatial risk factors

There is a limited literature on spatial aspects of TB transmission risk, although anecdotes abound.

> *'Forty or fifty years ago skin-test surveys in remoter areas of Africa showed few positives among the population. For instance in Rhodesia (now Zimbabwe and Zambia) a Colonial Service doctor told me in the 50s that the high positivity rates were all along the railways. The rate in the communities decreased to very low levels as he tested in districts far removed from it. In the 50s rates were also very low in hill areas of Nepal away from the established trade routes.'*
>
> *Crofton, 1993*(193)

Many surveys have mapped TST positivity on national or supranational levels, enabling, e.g., demonstration that the ARTI is higher in Nairobi than in rural parts of Kenya(194). However, typically, such surveys do not locate children more finely than to the region or, perhaps, the school attended. They may not collect detailed data on risk factors for MTB infection on each child.

Many putative risk factors for MTB infection – e.g. access to healthcare for adults with TB disease or adult HIV prevalence – exhibit variation on a fine spatial scale. Substantial variation in both access to healthcare and HIV prevalence have been

demonstrated between communities within the 20 x 20km Africa Centre surveillance area(169,170,195).

A small number of studies have looked at spatial variation in TST positivity at a finer resolution. A study in Puerto Rico(196) found that steep spatial gradients in TB notifications between adjacent communities were not mirrored by gradients in TST positivity, suggesting the gradient in notifications might be an artefact generated by differences in access to healthcare. The same study found higher rates of TST positivity in more rural areas.

In rural Karonga District in Northern Malawi, TST positivity is highest in the local town and on the lake shore than it is in a hilly region in the North of the study area(21). In the same community, Palwasha Khan has recently described an association between TST positivity in preschool children and maternal HIV positivity, the number of adults living in the home and whether a case of TB had been diagnosed in an adult living within 200 metres of their home(197). A similar finding has been reported in a township near Cape Town, where children living on the same residential plot as an adult case of TB, especially smear positive TB, were more likely to be TST positive (a plot might contain between 1 and 20 houses) (198).

There are a larger number of studies looking at spatial patterning in TB disease, usually using notification data. For example, Zelner(199) described spatial clusters of MDR-TB disease sharing a common 24 loci MIRU-VNTR strain type in Lima, Peru. In this study, active case finding was undertaken among the contacts of notified cases. Jenkins described pockets of higher MDR-TB notifications in both Georgia(200) and Moldova(201). In the Moldova study, high rates of both TB and MDR-TB were seen

in communities where more cases reported previous incarceration. However, such studies are unable to differentiate spatial variation in MTB transmission from spatial variation in risk factors for progression from infection to disease.

In this chapter, I test for micro-geographical clustering in TST positivity among children in the Africa Centre household surveillance programme. In the next chapter, I explore risk factors for TST positivity in the same children. Many of the risk factors I explore in that chapter are measured at the household and community level.

## Objectives

- To measure annual risk of MTB infection in a rural community in KwaZulu-Natal, South Africa.
- To describe spatial clustering of prevalent MTB infection in children aged six to eight years in this community.

## Methods

### Power calculations

Power calculations for my tuberculin school survey were based on the risk factor analysis, which is presented in the next chapter.

### Sampling frame

I planned to enrol young school-going children. This was to allow inference to be made about recent MTB transmission whilst having the logistical convenience of

being able to access children at school. I elected not to sample children in the reception year (ages 5-6 years) as testing without their parents present might be challenging.

I recruited children registered in the Africa Centre household surveillance programme. Given uncertainty about recruitment rates, I decided to attempt to enrol all Grade 1 and 2 students attending schools in the surveillance area. Therefore, the sampling frame was children born in 2005 or 2006. I estimated that approximately 3000 resident children in that age bracket were enrolled in local primary schools.

Forms were printed for all registered children, except those in a handful of 'avoided' households where visits could not be undertaken safely. However, fieldworkers were requested to only obtain consent for testing if children were attending Grade 1 or 2 in one of the 38 primary or lower primary schools in the surveillance area. Where consent was erroneously obtained for children attending Grade R or Grade 3 at one of the 38 schools, we elected to test as per their parent or guardian's wishes. Children reported to be currently receiving TB treatment were not eligible for testing.

### Residence and inclusion in analytical datasets

It is possible for children to be registered in the Africa Centre household surveillance programme but to be non-resident. The surveillance programme registers all individuals who are considered members of a household in the surveillance area, regardless of where they reside. It is possible that children with a household connection to the surveillance area might be resident on the periphery of the surveillance area, attend a school within the surveillance area, and that a parent or

guardian resident in the surveillance area might have signed a consent form enabling participation in this study.

As I wanted my estimates to apply to a distinct population and as several of my analyses required children to be located to a place of residence, I elected to restrict my analyses to children in the original sampling frame (see above), who were resident in the surveillance area on 25 June 2013, the day we commenced recruitment. For data on residency, I used the September 2015 release of the Africa Centre's 'Individual Residencies' table.

### Approaches to study recruitment

We planned to get written informed consent for testing from a parent or guardian in one of two ways. First, during a four monthly surveillance visit immediately preceding the tuberculin survey, fieldworkers attempted to get written informed consent if a parent or guardian were present. Where this was not possible (usually as a result of the parent or guardian not being present) and where there was no refusal to consent, letters explaining the study in isiZulu and English and consent forms were sent home from schools with children. These consent forms were pre-printed with the child's name and unique barcodes. When sending letters home from schools, the study team looked for children in the schools closest to where they were last recorded as residing.

### Identifying consented children

School visit forms with barcodes were printed for all children, unless there had been a refusal to consent. This enabled us to test children should we find valid signed consent forms had been returned to the school. To proceed to testing, the study

nurse required the child to provide three identifiers to confirm that they were the child we had parental consent to test. For this, a set of pre-printed identifiers appeared on the school visit form – child's name, child's date of birth, mother's name, the name of the head of household, the area in which their homestead was located, and the Bounded Structure ID. The Bounded Structure ID is a unique identifier used by the Africa Centre to identify homesteads. A piece of plastic with the number on it is often tacked onto doors or gateposts. It proved unhelpful in identifying small children who were unable to remember this number! Children were usually able to provide their name, the names of household members and the name of the local area in which they lived.

### Cold chain and tuberculin storage prior to use

Tuberculin was received from the suppliers with documentation of cold chain compliance and stored in a fridge in an access controlled area. Minimum and maximum temperatures were recorded daily using an analogue thermometer with adjustments to the thermostat made when necessary to achieve a constant temperature of 2-8 degrees Celsius.

For school visits, the vials of tuberculin were transported in cooler boxes, separated from direct contact with ice blocks by a layer of Styrofoam. Minimum and maximum temperatures during each visit were recorded using another analogue thermometer with the same target temperature.

Care was taken to prevent the tuberculin from freezing and avoid leaving tuberculin in direct sunlight. Once drawn up into syringes, the tuberculin was used or discarded

within 90 minutes. Once open, tuberculin vials were used or discarded within 24 hours.

### Tuberculin testing

The conduct of the survey was broadly in concordance with the KNCV generic protocol for tuberculin school surveys(202), except that we required written consent from parents and (a request from the ethics committee) written assent from the child.

In brief, testing was undertaken in a quiet area identified during a visit to the school prior to testing. Study equipment, including alcohol gel, disposable gloves, a sterilisable surface, sharps containers, and medical waste bins, were transported from the Africa Centre each morning. The study team comprised a research assistant, who fetched children from class, and a research nurse who managed the study paperwork and administered and read tests.

The study nurse administered a standardized dose of tuberculin – 2 Tuberculin Units in 0.1ml – to each child. The tuberculin was RT23/Tween 80 (Statens Serum Institut, Copenhagen, Denmark). The tuberculin was administered by intradermal injection into the volar aspect of the forearm using a 26 gauge needle. Skin was cleaned with an alcohol wipe prior to injection only if visibly dirty. Tests were not administered if the child had a febrile illness or skin disease.

We aimed to read TST reactions at 72 hours but permitted reading at 48 hours or 96 hours where necessary. Measurements of the extent of any induration were made having carefully palpated the margins of the reaction. The study nurse measured the

maximum transverse diameter in integer millimetres using a transparent ruler. After documenting the TST measurement, the study nurse inspected for a BCG scar.

I observed the first 200 TSTs administered and read by the study nurse and undertook regular school visits thereafter to oversee testing.

A small number of tests were administered or read at participants' homes. This was done to bolster recruitment rates in Indlovu Village, an area of Reconstruction and Development Programme (RDP) housing on the N2 national road opposite KwaMsane township. These home visits were necessary because of high absentee rates at school visits scheduled after the end of year exams.

Letters in isiZulu and English explaining the TST result were given to children to take home. Children with reactions of ≥10mm (or ≥5mm if HIV positive) were advised to attend Hlabisa Hospital for a clinical review if there was an additional risk factor for TB disease, such as recent TB contact, HIV positivity, signs or symptoms of TB, or failure to thrive. Parents were able to call a dedicated study phone for advice and we facilitated travel to Hlabisa Hospital by covering the cost of the return journey (for the outbound journey, there was a regular Government bus service to the hospital leaving from Primary Healthcare Clinics).

### Estimating the prevalence of MTB infection

I planned to use the TST cut point obtained using mixture analysis as my primary measure of TST positivity. I used the *normalmixEM* tool in R (R Foundation for Statistical Computing, Vienna, Austria) to fit a two component normal model to the distribution of non zero reactions observed(181,182).

I also undertook a sensitivity analysis to look at the impact of using alternative means of defining TST positivity.

≥ 10mm

≥ 14mm x 1.22

≥ 15mm

Reactions at the mode plus twice the proportion of reactions bigger than the mode

Reactions of 17mm plus twice the proportion of reactions bigger than 17mm

In Stata version 14.1 for Mac (Stata Corp, College Station, Texas, USA), I calculated proportion infected using each threshold, with a 95% confidence interval. I accounted for clustering by school using robust standard errors.

For each of the mirror methods, I used the same approach for reactions equal to or greater than the mode. I then repeated it for reactions greater than the mode. I summed these risks to obtain an estimate of the prevalence of infection. I summed the ends of the two sets of confidence intervals to obtain a confidence interval for MTB infection prevalence.

### Planned subgroup analyses

I calculated MTB infection prevalence for each stratum of the following variables.

- Age in years
- Sex
- Urban versus rural residence

- By quantiles of household asset ownership
- By whether the child had ever lived outside the surveillance area
- By whether the child was reported to have had prior contact with a person with TB

I used the same approach as for the overall estimates with the adjustment for clustering by school (robust standard errors) taking into consideration correlations in the entire dataset not only between children in the stratum of interest.

To classify children as urban or rural, I located them to the homestead in the surveillance area in which they were resident on 25 June 2013 and used standard Africa Centre definitions of urban and rural, which are based on population density.

To obtain wealth quantiles, I located children to this same homestead. Using data on every household in the surveillance area, I calculated a household wealth index by principal components analysis (PCA) (203). I included 2013 data on household ownership of the following assets – bed net, bed, bicycle, block maker, car, car battery, cattle, electric hotplate, electric kettle, fridge, gas cooker, Kombi/lorry/tractor, kitchen sink, motorcycle, other livestock, Primus stove, radio, sofa, sewing machine, table and chairs, telephone, cell phone, TV, video, wheelbarrow plus the fuel used for cooking and the sanitation facilities available. The use of this standard set of Africa Centre assets data and PCA allowed comparison with other studies undertaken at the centre, which have used the same approach to calculating household wealth.

Measures of socio-economic position based on household asset ownership are readily measured in household surveys. Household asset ownership is a medium to long term measure of wealth – which is what I wanted – and, unlike income, has

been shown to be relatively unresponsive to short term economic shocks.(204). The performance of asset indices are very dependent on the ability of the assets chosen to discriminate between richer and poorer households. Limited variation in patterns of asset ownership or the failure to include assets that distinguish between richer or poorer households can result in 'clumping' or 'truncation' of asset scores thus limited discriminatory power(203). Asset scores calculated using PCA are a relative rather than an absolute measure of socioeconomic position(203). They may not capture the function or quality of included assets but, if a sufficiently discriminatory set of assets are used, that additional detail may not be as important (205).

One potential alternative to a household wealth index would have been to look at the educational attainment of the mother or the household head. Educational attainment is another relatively stable measure of socioeconomic position, conceptualised as a measure of human capital as opposed to material wealth(206). However, in the communities around the Africa Centre, as a result of orphanhood and labour migration, many children live in households headed by a grandparent(207). Given educational opportunities for black people in South Africa changed dramatically after 1994 and high rates of maternal orphanhood(207), I did not think that maternal or household head educational attainment was likely to perform well as a measure of socioeconomic position.

### Estimating the annual risk of MTB infection from prevalence

To estimate ARTI from MTB prevalence, I needed mean age at testing. To obtain mean age at testing, I calculated age at testing for each child. Then, in Stata, I calculated mean age with a 95% confidence interval, adjusting for clustering by school using robust standard errors. With little variation in age in the sample and no

reason to think the distribution of ages would vary by school, I did not expect the values adjusted for clustering to differ substantially from the crude mean.

ARTI was calculated as per Nyboe(190).

$$ARTI = 1 - (1 - prevalence)^{1/mean\ age}$$

To obtain conservative confidence intervals on this annual risk, I repeated the calculations using the ends of the confidence intervals for prevalence and age that resulted in the widest confidence interval for the ARTI estimate. More complex methods, such as bootstrapping, were not attempted. With little variation in the ages of the children, I did not think bootstrapping would substantially affect my conclusions.

### Spatial analyses

The spatial distribution of positive and negative tuberculin tests was mapped on a point map with a random error introduced to protect confidentiality.

To assess for evidence of spatial heterogeneity, I used the Kuldorff Spatial Scanning Statistic(192) implemented in SaTScan version 9.3(208). In the primary analysis, children were located to the homestead in which they had lived for the longest period of time. The most likely clusters, obtained using the data without the random error, were then overlaid onto the point map with the random error.

In a sensitivity analyses, children were located to the homestead in which they were living at the start of the study. I also repeated the analysis restricted to children who

had never moved. I repeated the analysis including children reported to have received treatment for TB disease as cases (they were excluded from the previous analyses). I also undertook a sensitivity analysis varying the TST cut point used, as described above.

### Ethics approvals

Approval to undertake the tuberculin survey were obtained from the University of KwaZulu-Natal Biomedical Research Ethics Committee (BF075/13), the Africa Centre Community Advisory Board, the medical director at Hlabisa Hospital and the Department of Education (Hlabisa Circuit).

# Results

## Survey Participation

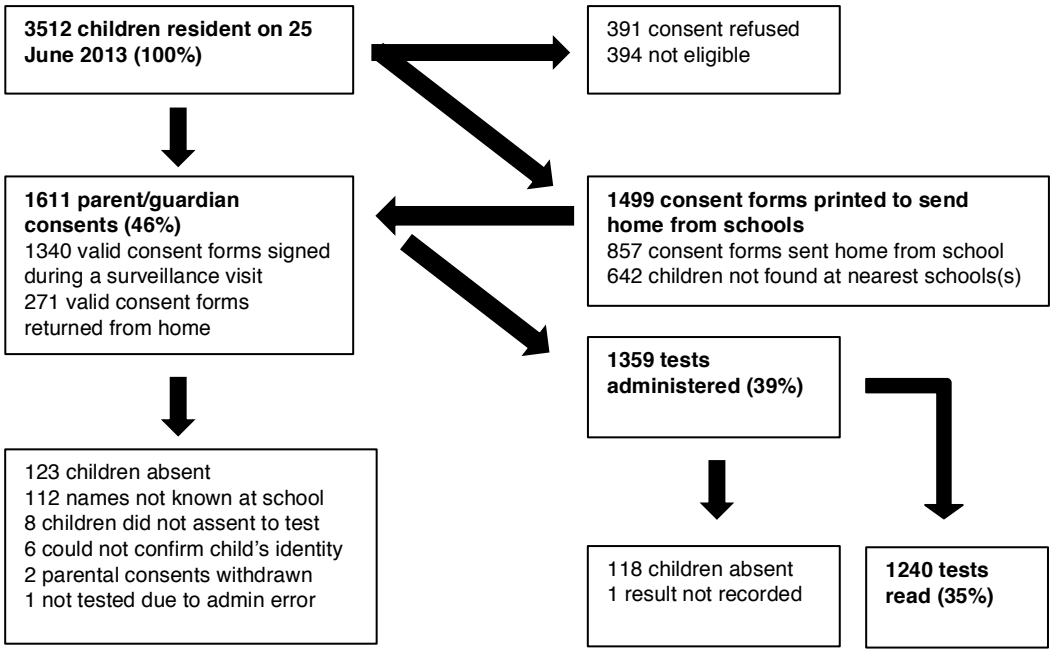



*Figure 11. Flowchart describing recruitment to the tuberculin school survey.*

Fieldworker visits with a view to obtaining consent were undertaken between 25 June and 27 October 2013. School visits were undertaken between 15 July and 5 December 2013.

In Figure 11, there is a flow chart describing recruitment to the study. In summary, 4479 children were included in the sampling frame of whom 3512 were resident at the start of recruitment. I obtained TST results on 1266 children of whom 1240 (97.9%) were resident at the start of recruitment.

The main barrier to recruitment was probably failure to make contact with a parent or guardian with only 391 children not participating as a result of documented parent/guardian refusal (11.1% of all resident children). However, the reasons for non participation following attempts to obtain consent using letters sent home from school are unclear – the majority of the consent forms distributed in this way were never seen again. The principal reason for non-eligibility was attending a school outside the surveillance area. The figures quoted for absenteeism do not include children we were able to test or tests that we were able to read by visiting the school on subsequent days.

We obtained test results on between 1 and 166 resident children at each of the 38 primary and lower primary schools in the surveillance area. These children were resident in 1100 households with between one and four children per household.

A comparison between eligible resident children and children for whom we have a TST result is presented in Table 3. Here, age is age on 25 June 2013. Household wealth and urban versus rural residence are based on the child's residence at the start of recruitment. The p values are Wald tests from a logistic regression model with a single categorical explanatory variable and success in obtaining a TST result as the outcome. This model did not account for clustering by school or household.

*Table 3. A comparison between children for whom I had a TST result and children for whom I did not.*

| | Number of children with TST result (column %) | Number of children without a TST result (column %) | p value |
|---|---|---|---|
| Overall | 1240 (100) | 2272 (100) | |
| Sex | | | |
| Male | 622 (50) | 1150 (51) | 0.80 |
| Female | 618 (50) | 1122 (49) | |
| Age in years [1] | | | |
| 6 | 339 (27) | 545 (24) | <0.0001 |
| 7 | 639 (52) | 1047 (46) | |
| 8 | 262 (21) | 680 (30) | |
| Residence | | | |
| Urban | 73 (6) | 171 (8) | <0.0001 |
| Peri-urban | 290 (24) | 814 (36) | |
| Rural | 868 (71) | 1260 (56) | |
| Household wealth | | | |
| Poorest quintile | 313 (26) | 357 (17) | <0.0001 |
| Second quintile | 299 (19) | 423 (20) | |
| Third quintile | 248 (21) | 445 (21) | |
| Fourth quintile | 245 (21) | 464 (22) | |
| Wealthiest quintile | 159 (13) | 418 (20) | |
| Residence outside surveillance area | | | |
| Never | 968 (78) | 1767 (78) | 0.84 |
| Ever | 272 (22) | 505 (22) | |
| BCG vaccine | | | |
| Vaccinated | 1022 (100) | 1762 (100) | 0.15 |
| Not vaccinated | 1 (0) | 8 (0) | |

*1. Age here is that on 25 June 2013, the day we began to consent children.*

We were slightly less likely to have TST results on children living in peri-urban versus rural areas, on eight years olds and on wealthier children. Whilst the p values were low, the absolute differences in participation rates were modest.

### The distribution of positive reactions

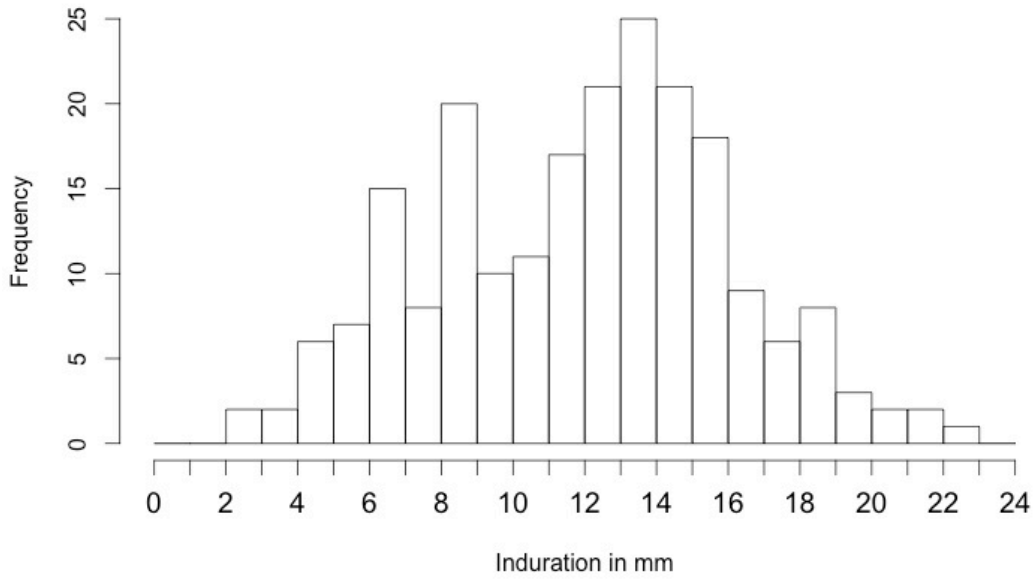



*Figure 12. A histogram showing the distribution of reaction sizes among the 214 children with non-zero TST reactions.*

The distribution of the 214 non-zero TST reactions is shown in Figure 12. There was a clear mode at 14mm, which is similar to the distributions seen in some surveys in the Western Cape of South Africa(12,185,209), though slightly lower than seen in surveys in a township near Cape Town(210,211). The distribution did not show the usual digit preference for multiples of five(212). The peaks at seven and nine were not accompanied by troughs at five or ten millimetres. This suggests that they were not explained by inverse digit preference resulting from deliberate avoidance of values that are known to be preferred.

### Mixture analysis

I performed mixture analysis, using *normalmixEM*. The model required two normally distributed components but applied no other constraints and used uninformative priors. The results proved unstable, predicting one of two distinct outcomes depending on the seed used (Figure 13). Most often, two separate normal distributions were predicted, one with a mean at 7.3mm and the second with a mean at 14.0mm. This result suggests that the standard cut point of ≥10mm would be most discriminatory. However, nearly as frequently, mixture analysis predicted normal distributions with means at 12.2mm and 14.6mm. This result suggests no distinct distribution of non-specific reactions – i.e. that all non-zero TST reactions represent MTB infection.

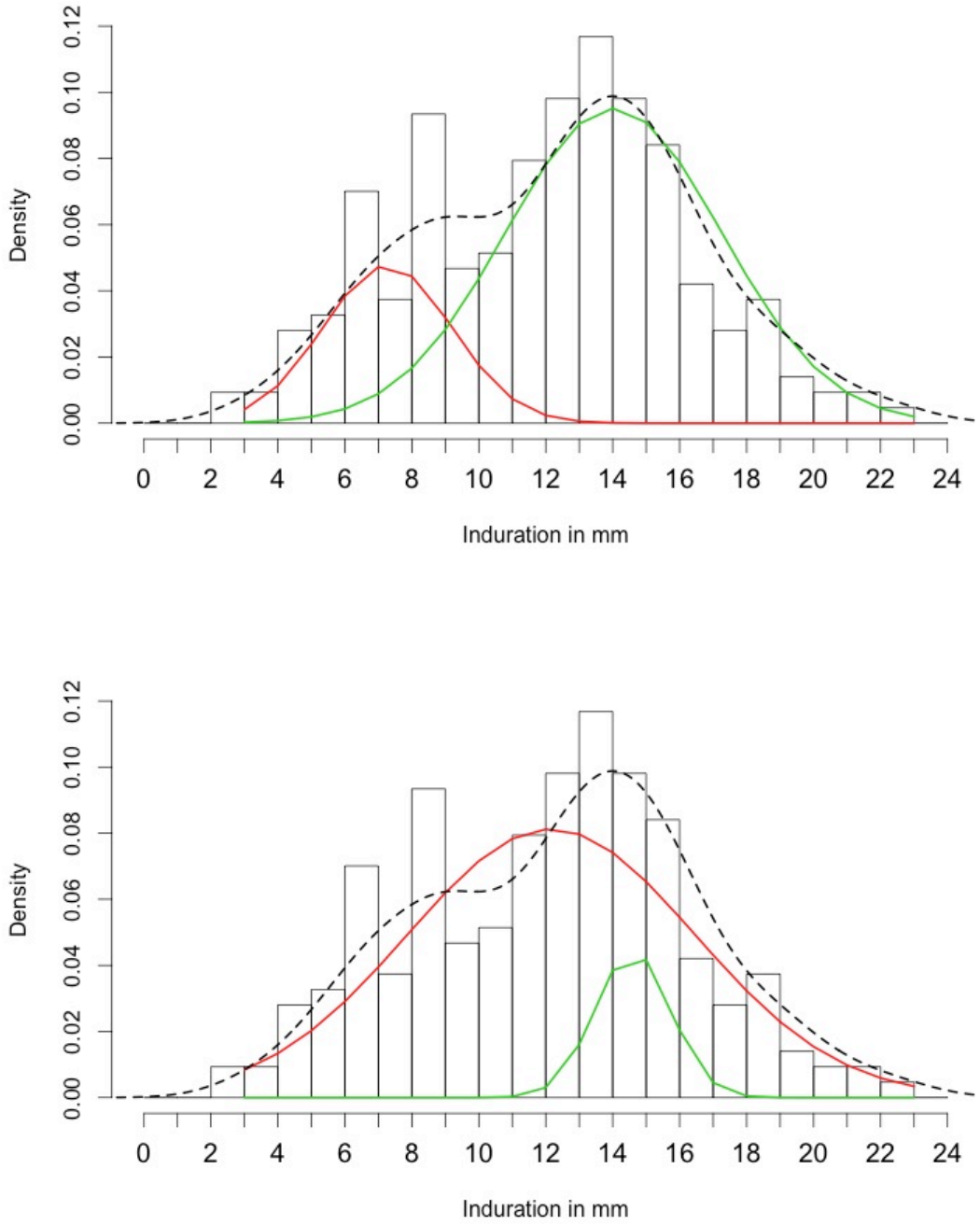



*Figure 13. Two alternative underlying component distributions, both predicted by mixture analysis, overlaid on the distribution of non-zero TST reaction sizes.*

### Generation of the wealth index

The 2013 household wealth index was generated using the full list of assets listed above, with toilet type and source of energy converted into sets of dummy binary variables. The first principal component explained 16.6% of total variance. The histogram of 2013 household asset scores (Appendix 4) showed some clumping at the top of the distribution, suggesting that the set of assets that were used might fail to discriminate between the wealthier households.

### Estimated prevalence of MTB infection

The prevalence of MTB infection overall and by subgroup is presented in Table 4. As expected, the definition of MTB infection used substantially affected prevalence estimates. The thresholds based on mixture analysis are highlighted. The expected higher prevalence of MTB infection in older children was not seen, probably because the age range was narrow and there were relatively few six year olds.

There were few children on whom we had a TST result (n=73) who were living in an urban area so those estimates should be treated with caution. Among children consented for TST testing, 23.9% reported a history of TB contact. In most cases (88.1%) this was a household TB contact. Given the limited data, I elected not to present separate estimates for children with a history of TB contact outside the household.

*Table 4. The prevalence of MTB infection overall and by subgroup using various definitions of TST positivity (%).* [1]

| | All positive reactions | 10mm | 14mm x 1.22 | 15mm | Mirror (mode 14mm) | Fixed mirror (at 17mm) | All positive reactions plus those on TB treatment | 10mm plus those on TB treatment |
|---|---|---|---|---|---|---|---|---|
| **Overall** | 17.3 (14.3-20.7) | 12.4 (10.2-15.0) | 9.3 (7.2-12.0) | 5.6 (4.1-7.7) | 13.3 (10.1-17.5) | 4.3 (2.6-7.1) | 19.1 (15.8-22.9) | 14.4 (11.8-17.4) |
| **Sex** | | | | | | | | |
| Male | 15.8 (11.8-20.7) | 11.4 (8.3-15.4) | 9.0 (6.2-13.0) | 5.5 (3.6-8.3) | 12.9 (8.6-19.0) | 4.5 (2.4-8.6) | 18.4 (13.9-23.8) | 14.2 (10.5-18.8) |
| Female | 18.8 (15.6-22.5) | 13.4 (10.9-16.4) | 9.7 (7.1-13.1) | 5.8 (3.8-8.7) | 13.8 (9.7-19.5) | 4.0 (2.1-7.7) | 19.8 (16.2-24.0) | 14.5 (11.6-18.0) |
| **Age in years** [2] | | | | | | | | |
| 6 | 18.1 (12.3-25.7) | 11.8 (7.3-18.6) | 5.9 (2.7-12.7) | 4.9 (2.2-10.4) | 9.7 (4.4-20.9) | 2.8 (0.7-11.2) | 17.5 (13.3-22.8) | 12.6 (9.0-17.4) |
| 7 | 16.7 (13.4-20.6) | 12.6 (10.0-15.6) | 11.2 (8.8-14.2) | 6.9 (5.2-9.3) | 16.1 (12.4-20.9) | 5.3 (3.3-8.6) | 18.9 (15.0-23.6) | 14.6 (11.6-18.3) |
| 8 | 17.9 (12.9-24.2) | 12.4 (8.9-17.0) | 7.6 (4.5-12.4) | 3.8 (2.0-7.0) | 10.0 (5.7-17.2) | 3.1 (1.3-7.4) | 21.5 (15.7-28.6) | 15.9 (12.0-20.8) |
| **Residence (2013)** | | | | | | | | |
| Urban | 28.8 (23.6-34.5) | 19.2 (13.6-26.3) | 15.0 (7.6-28.1) | 9.6 (5.1-17.4) | 21.9 (11.3-40.4) | 9.6 (4.6-19.5) | 30.7 (24.6-37.4) | 21.3 (14.8-29.8) |
| Peri-urban | 17.2 (11.6-24.8) | 14.1 (9.9-19.8) | 11.4 (8.1-15.8) | 7.2 (4.3-12.0) | 16.6 (10.9-24.9) | 4.8 (2.5-10.1) | 18.1 (11.8-26.7) | 15.0 (10.1-21.8) |
| Rural | 16.4 (13.8-19.3) | 11.3 (9.3-13.7) | 8.2 (6.1-10.8) | 4.7 (3.3-6.8) | 11.4 (8.2-15.6) | 3.5 (1.8-6.7) | 18.5 (15.7-21.7) | 13.6 (11.1-16.5) |
| **Household wealth (2013)** | | | | | | | | |
| Poorest quintile | 17.6 (13.1-23.1) | 12.1 (8.8-16.6) | 9.4 (5.9-14.7) | 5.4 (3.1-9.3) | 13.1 (7.9-21.4) | 4.8 (1.7-13.2) | 19.9 (15.5-25.1) | 14.6 (11.3-18.7) |
| Second quintile | 14.8 (11.3-19.3) | 12.2 (9.3-16.0) | 9.6 (6.8-13.3) | 4.8 (2.9-7.8) | 12.7 (8.6-18.7) | 3.9 (1.7-9.2) | 16.3 (12.6-20.9) | 13.7 (10.5-17.8) |
| Third quintile | 21.4 (16.2-27.7) | 16.1 (11.9-21.6) | 10.8 (6.8-16.9) | 5.6 (3.0-10.3) | 14.5 (8.6-24.1) | 5.2 (2.2-12.5) | 22.9 (17.2-29.9) | 17.8 (13.1-23.7) |
| Fourth quintile | 17.1 (12.1-23.7) | 10.6 (6.8-16.2) | 8.0 (4.2-14.6) | 5.3 (2.7-10.2) | 11.8 (6.2-22.2) | 2.9 (1.0-7.7) | 18.1 (12.9-24-9) | 11.7 (7.6-17.5) |
| Wealthiest quintile | 16.4 (10.8-23.9) | 11.9 (8.2-17.1) | 10.0 (6.6-14.9) | 8.2 (5.4-12.2) | 16.4 (10.8-24.5) | 5.7 (2.6-12.4) | 17.9 (12.5-25.0) | 13.6 (9.8-18.5) |
| **Residence outside surveillance area** | | | | | | | | |
| Never | 16.3 (13.2-20.0) | 11.9 (9.7-14.5) | 9.1 (6.9-11.8) | 5.7 (4.0-8.0) | 13.1 (9.7-17.7) | 4.1 (2.4-7.2) | 18.1 (14.5-22.3) | 13.8 (11.0-17.0) |
| Ever | 20.6 (15.7-26.6) | 14.3 (10.9-18.6) | 10.3 (6.9-15.3) | 5.5 (3.2-9.2) | 14.0 (8.9-21.7) | 4.8 (2.1-10.9) | 22.6 (17.8-28.2) | 16.5 (13.1-20.5) |
| **History of household TB contact** | | | | | | | | |
| No | 15.4 (12.4-19.0) | 11.1 (8.7-14.0) | 7.9 (5.9-10.6) | 4.7 (3.2-6.8) | 11.1 (8.0-15.5) | 3.8 (2.4-6.2) | 15.4 (12.4-19.0) | 11.1 (8.7-14.0) |
| Yes | 23.9 (18.9-29.8) | 17.5 (13.4-22.5) | 14.4 (10.6-19.3) | 8.9 (6.1-12.9) | 20.7 (14.8-28.7) | 6.1 (2.5-14.3) | 23.9 (18.9-29.8) | 17.5 (13.4-22.5) |

1. All estimates adjusted for clustering by school using robust standard errors.
2. Here age is age at testing except for calculations including children reported to be on TB treatment, where I used age on 25 June 2013.

### Estimated annual risk of MTB infection

The mean age on the day the TST test was administered in the children with a TST result was 7.71 years (95% CI 7.66 – 7.75 years), adjusted for clustering by school. ARTI estimates were as follows.

2.4% (95% CI 2.0-3.0%) including all positive reactions as cases

1.7% (95% CI 1.4-2.1%) using a threshold of ≥10mm

0.8% (95% CI 0.5-1.0%) using a threshold of ≥15mm

Children on TB treatment had the same mean age at enrolment as those who had a TST result (7.4 years). ARTI estimates including those children (as cases) and assuming they had the same age distribution as children with TST results were as follows.

2.7% (95% CI 2.2-3.3%) including all positive reactions and children reported to be receiving TB treatment as cases

2.0% (95% CI 1.6-2.5%) using a threshold of ≥10mm and including children reported to be receiving TB treatment as cases

We collected no additional data on the 'TB treatment' children were receiving. Furthermore, those children were slightly more likely to be included in the sample as there was not the same attrition as a result of failure to make contact with a parent or guardian, absenteeism, or test refusal. The estimates including self-report of TB treatment might therefore slightly overestimate ARTI.

### The spatial distribution of MTB infection

Spatial heterogeneity in MTB infection was not clearly apparent locating children to the homestead in which they were living on 25 June 2013 nor locating them to the homestead in which they had lived for the longest period of time. There was only a weak suggestion using a threshold of ≥15mm that an area just north of KwaMsane Township had a lower than expected number of MTB infections. The evidence for this was stronger locating children to the homestead in which they had lived for the longest period of time (p=0.06) than locating them to the homestead in which they were living on 25 June 2013 (p=0.08). In the former analysis, no MTB infections were seen in the 150 children resident in the cluster versus 8.47 expected (Figure 14).

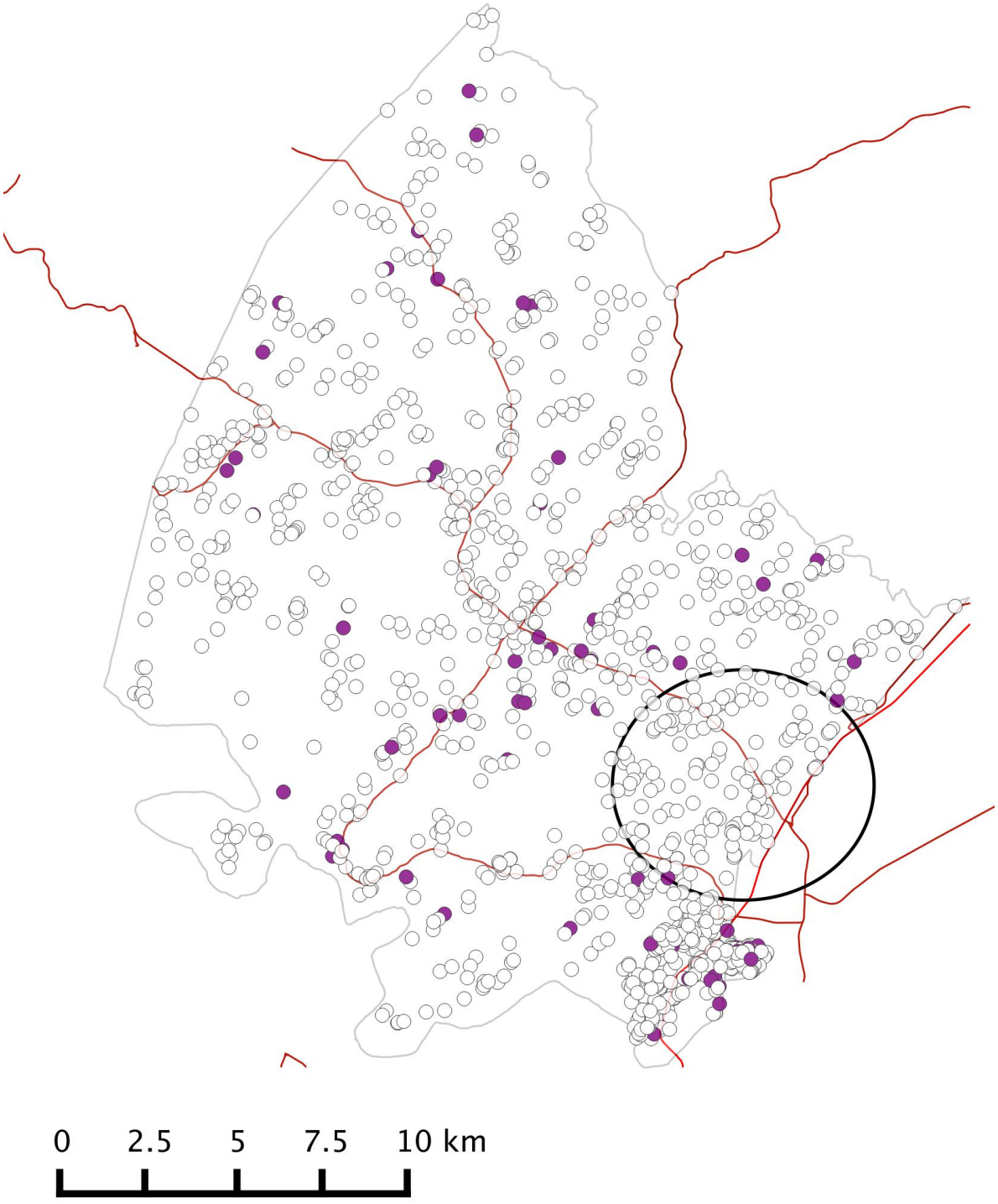



*Figure 14. A point map locating children to the homestead in which they lived for the longest period of time with children with TST reactions ≥15mm in purple.*

In the cluster, located to the north of KwaMsane Township, there were no positive reactions in 150 children whereas 8.47 were expected (p=0.06). A random error has been added to the point locations. Major roads are shown in red.

Better evidence for spatial heterogeneity in MTB infection was apparent in an analysis restricted to 583 children that been resident in the same homestead for most of their lives. Here, I opted to include children who had spent less than six months living elsewhere to maintain power.

In these analyses, there was evidence for an area of low infection prevalence in the same location north of KwaMsane classifying all positive reactions as MTB infection, both including (p=0.01) and not including (p=0.02) those reporting current receipt of TB treatment as cases. In the former analysis, no MTB infections were seen in 55 children resident in the cluster versus 9.77 expected (Figure 15).

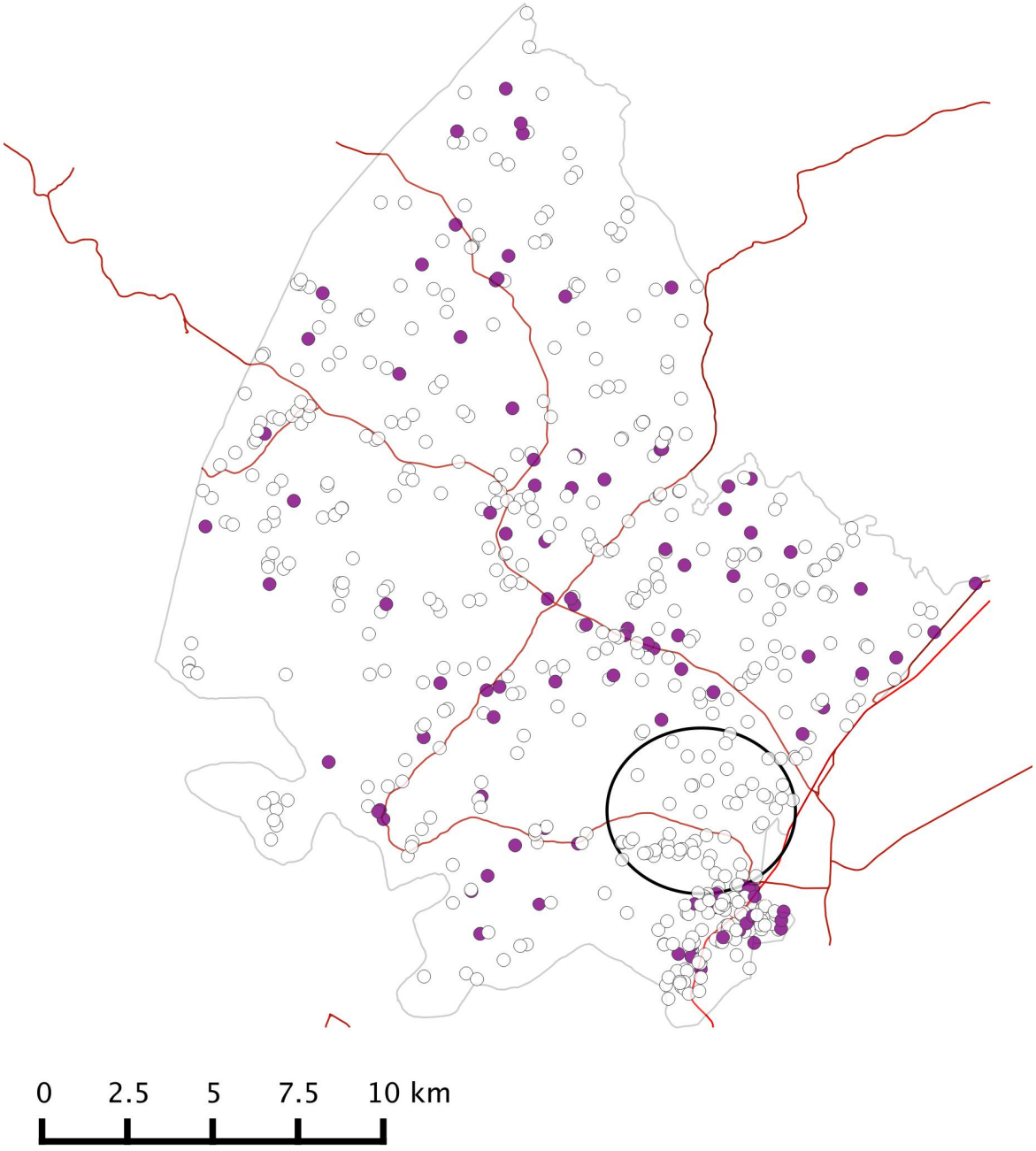



*Figure 15. A point map only including 592 children who had lived in the same homestead for most of their lives with children with non-zero TST reactions or reported receipt of TB treatment in purple.*

In the cluster north of KwaMsane Township, no MTB infections were seen in 55 resident children whereas 9.77 were expected (p=0.01). A random error was added to point locations.

There was also evidence for an area of high infection prevalence spanning parts of Indlovu Village and parts of KwaMsane Township. This was apparent using a ≥10mm threshold (p=0.05), a ≥15mm threshold (p=0.11) and using a ≥10mm threshold with individuals reporting current receipt of TB treatment also included as cases (p=0.11). Two of these clusters were very small, containing only four children. However, using the 15mm threshold, there were 8 cases in a population of 32 children, versus 1.59 expected (Figure 16).

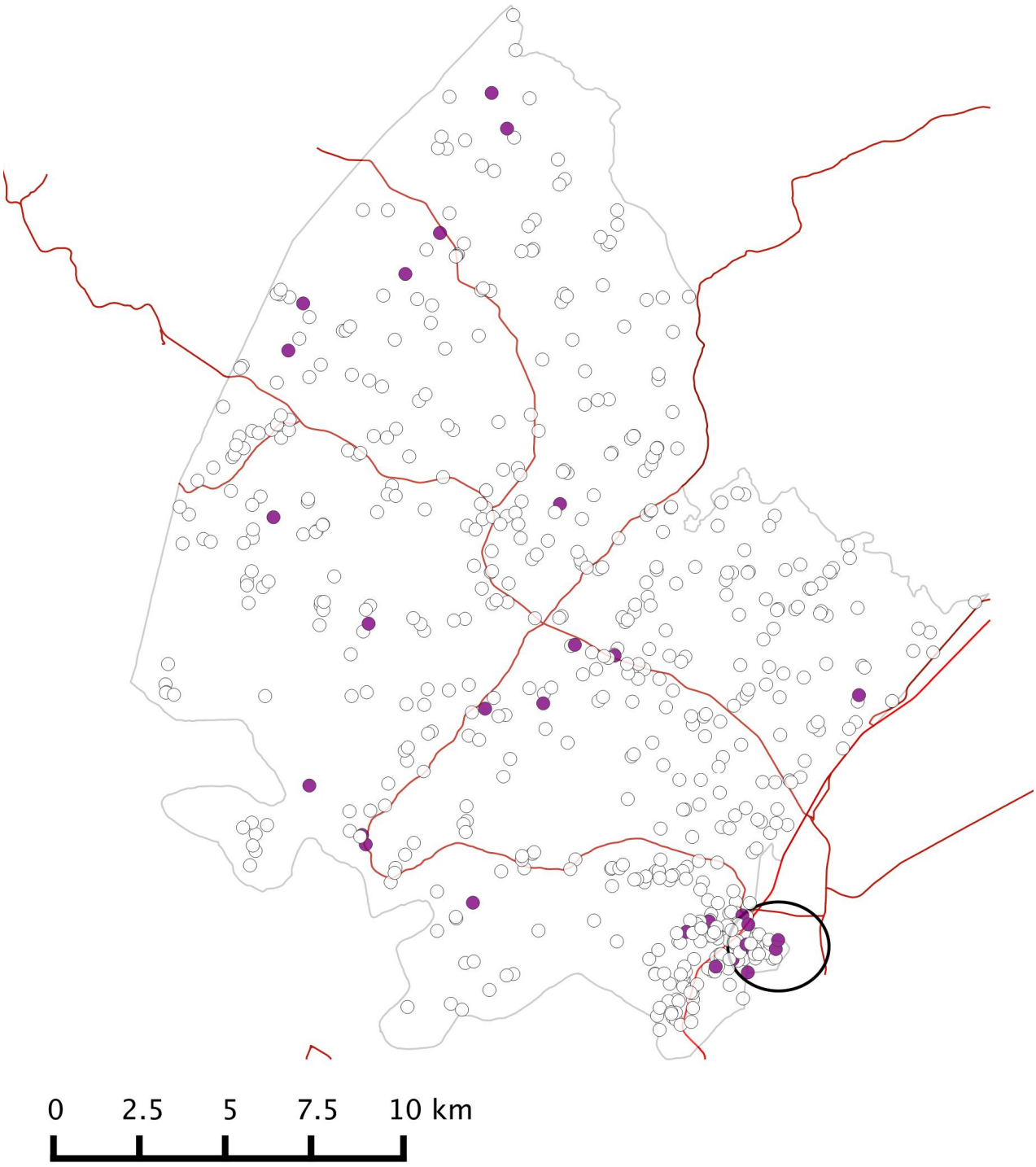



*Figure 16. A point map including 583 children for whom we had a TST result who had lived in the same homestead for most of their lives with children with TST reactions of ≥15mm in purple.*

In the cluster, located in KwaMsane and Indlovu Village, there were eight cases in a population of 32 children whereas only 1.59 were expected (p=0.11). Here, MTB infection was defined as a TST reactions of ≥15mm. A random error was added to the point locations.

Slight overlap between the high and low prevalence clusters probably explains the fact that in any single analysis we did not see both the high and low prevalence clusters.

## Discussion

I successfully obtained TST readings from 1240 children resident in the Africa Centre surveillance area. These children were broadly representative of the underlying population. To my knowledge this is the first TST survey undertaken in KwaZulu-Natal in forty years.

I demonstrated that, as expected, MTB infection was common and the ARTI high, though residual uncertainty about the appropriate TST cut point and the inherent limitations of the test, mean these figures should be taken as a proxy for rather than a precise estimate of the force of infection. There was some evidence for spatial heterogeneity in MTB infection with a small pocket of higher prevalence in KwaMsane Township and Indlovu Village and an area with lower prevalence immediately to the north of the township. This heterogeneity was only apparent using more specific cut points or restricting analysis to children who had mostly lived in the same household.

### Mixture analysis

That mixture analysis was unable to confidently ascribe a set of underlying distributions to the observed reaction sizes is consistent with the known limitations of the approach. It is expected to struggle where the frequency distribution is noisy as a

result of a small sample size or digit preference or where there is considerable overlap in the distributions(183). Here, whilst there was no evidence of digit preference, the sample size was relatively small and the mode was low, as has been seen elsewhere in South Africa(185,209).

The question regards which set of underlying distributions best fit these data is an interesting one.

The 1974 tuberculin survey in KwaZulu-Natal(96) found six percent of children had stronger reactions to an 'avian' purified protein derivative (presumably prepared from *Mycobacterium avium* or *Mycobacterium intracellulare*, two NTMs) than to tuberculin. However, the mixture analysis result with two distinct modes would suggest a much higher prevalence of non-specific reactions than that. It is possible that there was some residual BCG effect in these children who had almost all been vaccinated shortly after birth – BCG coverage was much lower in 1974(96). The near universal coverage of BCG vaccine precluded a comparison of reactions in BCG vaccinated and unvaccinated children. The alternative mixture analysis result, with no non-specific reactions, would be consistent with there being no 1-2mm reactions in my data. Such small reactions are common in settings where NTMs are prevalent(7).

### Infection prevalence and ARTI estimates

Our ARTI estimate of around 2% per year is high by African standards. For example, ARTI in Karonga, Northern Malawi, is approximately 1% per year and falling(151). Recent national tuberculin school surveys in Ghana(186) and Kenya(194) reported ARTIs of 0.0-0.6% and 1.1% per year respectively.

However, our results are comparable to other estimates from Southern Africa. For example, I modelled my analysis of the sensitivity of the ARTI estimates to the TST cut point chosen on a TST survey undertaken as part of the ZAMSTAR trial(185). In the Zambian communities in that survey, the ARTI estimates fell between 0.8% and 2.8% respectively(185), strikingly similar to our estimates from KwaZulu-Natal. Our data are also not too different to the 1974 ARTI estimate from KwaZulu-Natal of 1.4% per year. This is interesting given the huge changes that have occurred in the province, including substantial improvements in access to healthcare, the arrival of HIV and the roll out of antiretroviral therapy, a period of communal violence in the early 1990s, and the lifting of Apartheid era restrictions on freedom of movement.

TST surveys undertaken in the Western Cape of South Africa, however, have consistently estimated ARTIs that are higher than seen elsewhere in Southern Africa. For example, recent surveys from the Western Cape have estimated ARTIs of 2.5-4.2% per year in primary school children across eight different communities(185), ARTI of 4.1% in 5-17 year olds and 4.2% in 13-22 years olds in a township near Cape Town, and (analysing IGRA data using the same cross sectional approach as in the other studies) 7.3% per year in 12-18 year olds in the Cape Winelands. As outlined in Chapter 1, I am not aware of any definitive explanation for the very high force of infection seen in the Western Cape.

### Spatial heterogeneity

That, in a mobile population, spatial heterogeneity in MTB infection prevalence was only apparent when the analysis was restricted to children who had spent most of their lives resident in one homestead is an important methodological lesson.

The area of low MTB prevalence immediately north of KwaMsane Township is interesting. The evidence for this cluster was reasonably good. It was noted using the ≥15mm threshold in the full sample either locating children to the homestead in which they were living on 25 June 2013 (p=0.08) or to the homestead in which they lived longest (p=0.06). There was moderate evidence for this cluster in children who had mostly lived in the same homestead by two different definitions of MTB infection – all positive reactions (p=0.02) and all positive reactions or reported receipt of TB treatment (p=0.01). That it was not seen using the other thresholds is probably because of overlap with the high prevalence cluster immediately to the south. This 'low prevalence' community is neither remote nor clearly urban. Residents would be expected to have an intermediate HIV prevalence(195) and neither very ready access to healthcare and other amenities nor long journey times to reach such facilities(169,170).

There was less good evidence for the smaller area of higher infection prevalence in KwaMsane Township and Indlovu Village. This is a densely populated area with very high HIV prevalence, though ready access to healthcare and other amenities. The clusters here were only noted in the analysis restricted to children who had mostly lived in the same homestead. Furthermore, only using the ≥15mm threshold did this cluster affect more than a handful of children (n=32 rather than n=4). However, SaTScan would not have evaluated clusters with lower likelihoods that overlapped with these clusters. It is possible that, for this reason, I did not see other clusters of high infection prevalence in the same region. This phenomenon probably also explains why, in no single analysis, did I observe both high and low prevalence clusters.

## Limitations

### *Limited power*

I did not obtain TST results on as many children as I intended to. Whilst the confidence intervals around my ARTI estimates were acceptable, there was considerable uncertainty on the MTB prevalence estimates for some subgroups. There was also only weak statistical evidence to support the existence of high and low prevalence spatial clusters.

Power in spatial analysis is not simply a function of numbers but also of their spatial distribution. That said, I recruited better in rural areas where the incremental benefit per child tested will have been higher than in urban areas.

### *No formal validation*

With hindsight, I regret not including formal validation of the TST results documented by my study nurse. Having all tests administered and read by a single clinician was valuable in obtaining consistent and comparable results and I had no concerns about her technique. However, quantifying agreement in TST readings with blind double reading of a random sample of test reactions would have been reassuring.

### *No data on HIV status*

I elected not to collect data on these children's HIV status. Among the 2005-6 birth cohort in the Africa Centre surveillance programme, HIV prevalence is estimated to be less than 5% (personal communication, Dr James Ndirangu). This assumes an antenatal HIV prevalence of approximately 40%, a vertical transmission rate at that time of approximately 15% and substantial early mortality. Therefore, in

approximately one in twenty children in my study, the TST may have had reduced sensitivity(15). This may have led to an underestimation of the prevalence of MTB infection and of the ARTI, particularly if HIV positive children had greater exposure to MTB than other children.

### *Differential survival*

Of the 5690 children registered in the surveillance programme who were born in 2005-6, 281 (or 4.9%) had died by 25 June 2013 (Africa Centre data). By verbal autopsy, the major causes of death in young children in the community were acute respiratory infections, HIV/AIDS and birth asphyxia(173). If MTB infected children had a higher mortality than those not infected with MTB, such differences in survival could also result in me underestimating the ARTI. However, it is not usual to account for such differences when analysing TST survey data and thus my estimates may still be comparable with those from similar communities.

### *Assumptions*

A further reason I may have underestimated MTB transmission is that I used cross sectional rather than longitudinal data. Fine noted, 'Conventional cross-sectional analysis methods give an accurate estimate of incidence only if there is in fact no reversion, and if there are equal numbers of false negative and false positive tests.' (21) TST reversions have been documented to occur in longitudinal data (12), even applying stringent definitions of test conversion(21). Such longitudinal studies therefore tend to produce substantially higher ARTI estimates than do cross sectional studies. Therefore, whilst my data provide an ARTI estimate that allows comparison with other similar cross sectional studies, I may have substantially underestimated the force of infection. For example, in the study in Cape Winelands adolescents, a

7.3% per year ARTI estimate was obtained treating the data cross-sectionally versus a 13.0% per year ARTI obtained from longitudinal data(12).

I also have no data to support the stable incidence assumption. There has been no clear trend towards higher or lower TB notifications in the district(89). However, notifications are a complicated function of force of infection, rates of progression to disease, and the capacity of the healthcare system. If ARTI were not stable, our estimate would best reflect the force of infection midway between these children being born and being tested (i.e. approximately 2009).

Finally, caution needs to be exercised in applying force of infection measured in young children to other individuals in the community. Age assortative mixing is seen in many communities(157), including South African communities(70,71). Furthermore, infections in children can provide only indirect estimates of transmission in settings that children do not frequent (i.e. only if they are infected by an adult who had been infected in that space).

## Conclusions

I have demonstrated a high force of infection in this community in Northern KwaZulu-Natal and some evidence of spatial heterogeneity in infection prevalence. I have demonstrated the limitations of the mixture analysis approach, particularly in smaller datasets. My data have also suggested that population mobility can obscure underlying spatial patterns in the prevalence of longstanding conditions such as MTB infection.

Spatial clustering of MTB infections might reflect transmission occurring in that community or, alternatively, spatial clustering of risk factors for MTB infection. In the next chapter, I explore risk factors for MTB infection in these data.

# 4. Individual, household and community level risk factors for *M. tuberculosis* infection in the Africa Centre Demographic Surveillance Area

## Background

Despite its limitations, there is a long tradition of using the tuberculin skin test (TST) to understand MTB transmission with clear and expected gradients in TST positivity seen between higher and lower burden settings(213), over time as TB burden falls(214,215), according to characteristics of the index case(37,65–67) and according to the extent and duration of contact(35).

In this chapter, I seek to test a series of hypotheses about MTB transmission in a rural community in northern KwaZulu-Natal by linking the tuberculin school survey data to data from the Africa Centre household surveillance programme.

### Social contact

As outlined in the introduction, there is a growing body of evidence associating exposure to indoor public spaces with MTB transmission. However, with the exception of a small number of studies demonstrating an increased risk of TB infection or TB disease among individuals commuting by public transport(78,216,217), empirical studies have focussed on elevated risk in individuals with occupational exposure to public spaces(79–83).

Other than studies among miners, healthcare workers and prisoners, the relationship between exposure to specific indoor congregate settings and MTB infection in high burden settings in Southern Africa is not known.

A qualitative study in a Cape Town township(218) attempted to identify congregate settings that might play a role in MTB transmission based on the number of people present, duration of exposure and factors such as whether the space appeared well ventilated. The researchers suggested drinking establishments (including informal 'shebeens'), clinics and churches (including church groups that met in private homes) might be important sites of MTB transmission.

Also working in a Cape Town township, Robin Wood's group combined data on social contact patterns(70), data on ventilation obtained using $CO_2$ metres and data on MTB infection by age(219). Using Rudnick and Milton's adaptation of the Wells-Riley Equation(160), they were able to predict the spaces that might be important for MTB transmission in the community(60). The model suggested that 84% of MTB was acquired outside one's own home, that school and workplace were the main site of transmission for children and adults respectively, and that public transport and households might be important places in which MTB transmission between age groups occurred. The risks associated with schools and workplaces were driven by the length of time spent in these spaces, poor ventilation in schools, and the large number of potentially infectious adults present in workplaces. Important limitations to this work include the lack of location specific TB prevalence data and the strong assumptions of the model (full air mixing and no heterogeneity in infectiousness or susceptibility to infection).

A social contact pattern study has attempted to estimate the number of 'contact hours' (a product of the number of people in the space and the duration of exposure) attributable to specific spaces in various communities in Zambia and the Western Cape of South Africa(72). In the Western Cape, most contact hours occurred in workplaces whereas, in Zambia, most contact hours took place in churches. The study did not collect data on location specific TB prevalence or attempt to measure ventilation in these spaces.

Palwasha Khan recently reported data on incident MTB infection in under fives in Karonga, in Northern Malawi(220). Data were available on 1242 children who had both a baseline and a repeat TST and whose parents or guardians answered a set of questions about use of public space. After adjusting for age, sex, household socioeconomic position and population density, the odds of MTB infection were 3.6 (95% CI 1.2-11.1) fold higher in children who had attended church more than four times in the preceding year as compared to other children. The adjusted odds of MTB infection were also higher in children who had been to a healthcare facility at least once in the preceding year (aOR 3.1; 95% CI 1.0-9.8) and in children who travelled by minibus (aOR 1.8; 95% CI 0.9-3.6) but not in children with greater exposure to outdoor congregate settings such as markets and funerals.

### The prevalence of untreated adults

As outlined in Chapter 1, individuals with pulmonary TB, initiated on effective TB treatment, rapidly become non infectious. A major focus of TB control programmes is, therefore, finding TB cases and ensuring they are effectively treated(221). Prompt treatment reduces the duration of infectiousness and therefore reduces the prevalence of infectious cases. As noted in the previous chapter, in the community

living near the Africa Centre, distance to the nearest primary healthcare clinic is a major determinant of healthcare usage (169,170).

### HIV prevalence

The arrival of HIV in Southern Africa in the early 1990s resulted in dramatic increases in TB burden. However, strikingly, it was noted (in gold miners) that, over the 1990s, steep increases in TB incidence in HIV positive individuals were not associated with changes in HIV negative individuals in whom incidence was stable over the same period(222). This may suggest that HIV positive TB cases are not important sources of MTB infection.

The force of infection in gold mines is very high and an alternative explanation is that, in such settings, host susceptibility is the primary determinant of TB incidence, not transmission. However, as outlined in Chapter 1, there are reasons to think that HIV positive people might be less infectious than HIV negative people with TB. They progress faster to death or treatment(53,54). They have more smear negative or extra-pulmonary disease and less cavitary disease(68,69). Greater morbidity may reduce levels of social contact.

Much of this data comes from the era before widespread availability of anti-retroviral therapy in Southern Africa. ART clearly extends life expectancy(94,171) and there is a suggestion that individuals on ART have TB disease with a more HIV negative phenotype(68,69). However, individuals in ART care might have their TB diagnosed faster and ART reduces incident TB disease in HIV positive people by approximately two-thirds (although incidence after immune recovery is still higher than in HIV negative individuals) (223). Modelling suggests that the net impact of the ART roll out

on TB incidence is uncertain and critically dependent on the extent of immune recovery and the success of HIV prevention efforts(224). The net impact of ART and ART care on MTB transmission is, I would argue, also uncertain, depending on both the impact on TB incidence and the impact on infectiousness.

### Socioeconomic position

TB has long been considered the archetypal disease of poverty with social gradients in the prevalence of TB disease observed in Bangladesh(225–227), the Philippines(227), Vietnam(227), Kenya(227), Zimbabwe (through this did not reach statistical significance)(228), Myanmar (though the association was not seen after stratifying by rurality)(229), Zambia and the Western Cape of South Africa(230). Social gradients in TB prevalence might occur as a result of differences in the incidence of MTB infection, in progression from infection to disease or in disease duration.

There are fewer studies of the association between MTB infection and socioeconomic position and results have not been consistent. A large study in adolescents in the Western Cape found MTB infection associated with low parental income and low parental educational attainment(209). However, a study in Zambia found MTB infection to be associated with both household crowding and higher household wealth(231). The authors of the latter study speculate that the association they observed may be because wealthier households have better access to 'facilities' (congregate settings) in which transmission occurs or alternatively that less well constructed homes might be better ventilated. An association between poor household construction and reductions in TB disease (75,158), exposure to 'shared

air' (159) and modelled risk of MTB infection (232) have been observed in a range of different settings in Sub Saharan Africa.

## Objectives

1. To ascertain individual, household and area level risk factors for TB infection in this community, including the association between TB infection and exposure to indoor public spaces.

2. Specifically, to test the hypotheses that MTB infection in children is associated with:

- Social contact in homes, in public spaces, and in clinical spaces
- HIV prevalence in adults
- Low socio-economic position
- Poor access to TB treatment

## Methods

### Power calculation

To achieve 90% power to detect a 10% absolute difference between children exposed and children unexposed to a risk factor in the proportion of children with a positive test, I estimated that TST results on 1350 children would be needed (alpha=0.05, two tailed test). This assumed 20% would be TST positive overall (based on recent TST survey data from the Western Cape(185,209–211)), 50% exposed to the risk factor and a design effect of 2 (given sampling would be undertaken in schools). Previous studies of schistosomiasis infection in the same schools (unpublished) suggested such numbers were realistic and would provide sufficient power to conduct spatial analyses.

### Population

The population studied in this chapter is the same as that studied in the previous chapter – namely children aged 6-8 years, registered in the Africa Centre surveillance programme, resident at the start of enrolment, and attending one of the 38 primary or lower primary schools in the area.

### Outcome measures

Here, I used the thresholds to define MTB infection obtained by mixture analysis – all positive tests and TST ≥10mm. I also tested whether the key findings were robust in further sensitivity analyses: including children on TB treatment as cases or using a potentially more specific ≥15mm threshold.

### Locating children

For these analyses, I located children to the homestead in the surveillance area in which they were resident for the longest period of time.

## Conceptual framework

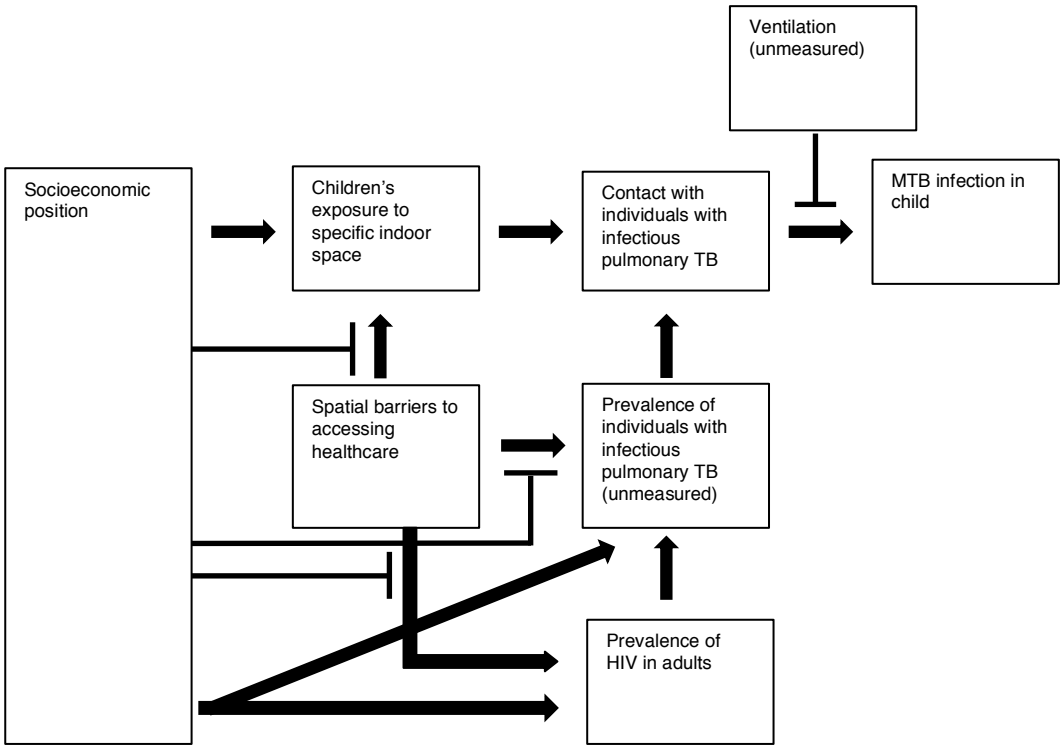



*Figure 17. Conceptual framework for the risk factor analysis.*

All analyses were adjusted for the child's age and sex.

I considered the prevalence of individuals with infectious pulmonary TB (not directly measured) and children's exposure to specific indoor congregate settings to be the immediate determinants of a child's contact with individuals with infectious pulmonary TB.

I thought socioeconomic position was a potential determinant of the indoor spaces children would be exposed to. For example, children in households who owned a car would be less exposed to public transport. The same might be true if there were no money for public transport.

I knew that distance was associated with attendance at primary healthcare clinics in the community(169,170) and thus might determine children's exposure to clinical spaces and also the prevalence of both untreated pulmonary TB and HIV(94,170) in adults. I thought that socioeconomic position might modify these effects given money for public transport or access to private transport might facilitate access despite spatial barriers. TB patients are known to struggle with the high costs of transportation in this community(233). Social gradients in access to TB treatment have been well described in other settings(225,229,234).

Finally, I knew that social gradients in prevalent TB disease were observed in many settings and therefore considered that socioeconomic position might directly affect TB prevalence via mechanisms other than determining access to healthcare. For example, insufficient dietary protein has been associated with prevalent TB disease in Zambia(230).

Social patterning in HIV prevalence has also been observed though this has been seen to change over time(235) and may be context specific. HIV is associated with TB prevalence, though to a lesser extent than it is associated with TB incidence(53,54). Socioeconomic position might, therefore, be associated with TB prevalence by determining risk of HIV.

In this framework, spatial barriers to accessing healthcare and socioeconomic position were upstream determinants of MTB infection in children, with their effects mediated, at least in part, through social contact patterns, adult HIV prevalence and the prevalence of pulmonary TB. Therefore, I elected not to adjust for social contact patterns or HIV prevalence when estimating their associations with MTB infection in

children. For the reasons outlined above, I planned *a priori* to test for an interaction between spatial barriers to healthcare and socioeconomic position. I repeated these analyses adjusting for population density in an attempt to obtain estimates of association between MTB infection in the child and distance to healthcare that were not explained by the more urban environments in which healthcare facilities tend to be situated.

I considered socioeconomic position and HIV prevalence potential confounders of the association between social contact patterns and MTB infection in children. I considered distance to healthcare a potential confounder of the association between exposure to clinical spaces and MTB infection (again exploring for potential interaction between socioeconomic position and distance to healthcare). Again, I repeated these analyses adjusting for population density in an attempt to isolate the effects of social contact in specific indoor spaces from other aspects of the more urban environments in which such spaces tend to be located.

I viewed socioeconomic position and spatial barriers to healthcare potential confounders of the association between HIV prevalence and MTB infection in children (and that there might be an interaction between socioeconomic position and distance to healthcare). I didn't adjust for children's social contact patterns as I thought it unlikely that there would be an association between HIV prevalence and children's social contact patterns after accounting for socioeconomic position and spatial barriers to healthcare. I repeated these analyses adjusting for population density as HIV prevalence is known to be higher in more urban parts of the surveillance area(195).

Where I found associations between putative risk factors and MTB infection in children, I repeated the analysis with history of household TB contact included in the model. This was to assess the extent to which association between the risk factor and MTB infection in the child might be mediated through their being exposed to individuals with TB disease in the household.

### Accounting for the multi-level structure of the data

Some variables were measured at the level of the child and some at the level of the household. Given the mounting evidence that much MTB transmission occurs outside the home (see Chapters 1 and 2), I was also interested in exploring the impact of risk factors such as deprivation and HIV prevalence at the community level. I therefore had a three level data structure to consider and also needed to account for clustering at the level of the school, given that is where testing was performed.

Given a four level model was not feasible, I needed another approach. For each of the community level variables, I was provided with a set of Gaussian kernels (236) centred on each homestead. These were generated by superimposing data at an individual or household level (e.g. HIV prevalence or household ownership of a specific asset) onto a 30 x 30 metre raster grid. A kernel of diameter 3km was then passed over each square in the grid and a Gaussian weighted prevalence estimate for each variable allocated to each cell. These kernels meant I could assign a unique value for HIV prevalence or area level deprivation to the 3 km radius immediately surrounding each household.

The approach had two major advantages. First, it allowed me to fit a three level model with variables at the level of the child and the household whilst also accounting

for clustering by school. Second, it avoided the modifiable areal unit problem (237) whereby choices regards the position of boundaries when generating community level data can substantially affect the inferences made.

### Measures of social contact

I added a module to the Africa Centre surveillance programme which was inserted into fieldworker bundles during surveillance rounds two and three of 2013. A separate form was printed for each eligible child. A key informant was asked to answer the following questions regards the child.

*For these questions, inside means within a space with a roof and at least three walls. If you are not sure the exact answer to a question, please estimate if you can...*

*1. In the last seven days, how many times has [name of child] been inside an indoor shop, an indoor market or supermarket, a bank or a post office?*

*2. In the last seven days, how many times has [name of child] been inside an indoor restaurant, takeaway, bar or shebeen?*

*3. In the last seven days, how many times has [name of child] used a bus or a minibus taxi? (count return journeys separately)*

*4. How many of these journeys lasted more than one hour?*

*5. In the last seven days, how many times has [name of child] been to a church service? (don't include outdoor services)*

*6. If they have been to a service, which church in which place did they attend most recently?*

*7. In the last month, how many times has [name of child] been inside a clinic or hospital, including if accompanying someone else?*

*8. In the last seven days, how many visitors to your homestead have been inside your household?*

*9. In the last seven days, how many households on other homesteads has [name of child] been inside?*

*10. Is [name of child] currently taking treatment for TB?*

To improve the accuracy of recall, I only asked about visits in the preceding week (or month for visits to clinic or hospital). Shebeens are informal drinking establishments. The stipulation regards outdoor services was included as the Nazareth Baptist Church – whose followers are colloquially referred to as 'Shembe' – is popular in the region. The church holds its services outdoors. In data cleaning, I set the number of church visits to zero where the name of the last church attended was a Nazareth Baptist Church.

Given concerns that models might become over parameterised, I planned to generate aggregate measures of the number of contacts in household spaces, non clinical public spaces and clinical public spaces, but only if the effect estimates for variables I planned to combine were not in opposite directions in unadjusted analyses. I planned also to look at the impact of household size (number of members per household) as a proxy for contact with other members of the same household. The household size variable was the number of residents on 1 June 2009 in the homestead in which the child lived longest. Where the child was not resident in that homestead on 1 June 2009, I substituted the number of residents in the homestead in which they were resident on 25 June 2013.

I did not generate an aggregate community level measure of social contact because I believed that adult-adult and child-adult contacts were the relevant community level measure. I had no data on these contacts.

### Quantifying socioeconomic position

Household asset indices were calculated by PCA, as described in chapter 3, using 2009 data.

This had two advantages. First, the distribution of household wealth indices have become increasingly skewed over time, presumably as goods have become more affordable (Appendix 4). Thus newer data might be less discriminatory. Second, 2009 is approximately midway through these children's lifetimes and, therefore, the 2009 data arguably best capture the socioeconomic position of these children over the period in which they were infected/at risk of infection regardless of whether that was recently or shortly after birth.

The area level deprivation score was derived by calculating a 'prevalence' of ownership of specific assets in households using a set of 3km kernels centred on each homestead. A similar (though not identical) set of assets were used – here, access to piped water and electricity were included (markers of socioeconomic position at the community but not necessarily the household level). The prevalence of ownership of each asset in the area around each household was then fed into a PCA. An asset score for each household was derived from the first principal component.

### Quantifying access to healthcare

Here, I used travel time in minutes from each homestead to the nearest PHC. These data were provided by my supervisor, Frank Tanser. They were derived using a published model that was validated against reported travel times(169).

### Quantifying HIV prevalence

The HIV data from the Africa Centre household surveillance programme is a rich resource. However, analysis is complicated by missing data. Between 26 and 46% of adults agree to provide a sample for HIV testing each year (238). More complete estimates of HIV status can be obtained by treating the data longitudinally with 60.6% and 68.4% of adults having tested at least once after three and five years respectively(238).

To obtain estimates of household HIV prevalence, I took the 2009 data. If an individual had no HIV test result from 2009, I used the result from 2010. If there was no result from 2010, I used the result from 2008. I then continued iteratively until either I found a non-missing result or reached 2005 (when our first children were born) or 2013 (when the children were tested).

Data on community HIV prevalence were from 3km Gaussian kernels provided by Frank Tanser. Detail about how they were derived can be found in the literature(195).

### Statistical approach

I began with a descriptive analysis describing the distribution of the exposures of interest in children for whom we had a TST result and comparing this to the population of resident children who were eligible to participate in this study. I then

described the crude associations between each exposure and MTB infection in the child.

Complementary log log regression is traditionally used to analyse data from TST surveys because coefficients can be interpreted as rate ratios which are more intuitive than the prevalence odds ratios obtained using logistic regression(239). However, in practice, where prevalence of the outcome is less than 0.2, coefficients will be almost identical and at prevalences of 0.2-0.4 they will be very similar(239). Here, because I struggled to get complementary log log regression models to fit to the data, I used logistic regression. I fitted mixed effects models to account for the multi-level structure of the data, with random intercepts for both household and school. I ensured all estimates were stable when the number of integration points was increased from 7 to 15.

### Tools

All analyses in this chapter were undertaken in Stata version 14.1 for Mac (Stata Corp, College Station, Texas, USA).

### Ethical approvals

The TST survey was approved by the University of KwaZulu-Natal Biomedical Research Ethics Committee (BF075/13). Additionally, the same committee approved an amendment to the ethics approval for the Africa Centre demographic and health surveillance programme (E009/00). This amendment permitted us to ask the social contact pattern questions and to merge the TST data with data from the surveillance programme.

## Results

### Generation of the wealth index

The household wealth index was generated as in chapter 3 but using the data from 2009. The first principal component explained 15.5% of total variance. The histogram of the 2009 household wealth scores showed less evidence of truncation than that describing the distribution of 2013 household wealth scores (Appendix 4). This suggests that this set of assets were more discriminatory in 2009 than in 2013, perhaps as a result of household goods becoming more affordable over time.

### The distribution of risk factors in my sample and in the wider population

Table 5 shows the distribution of risk factors in all eligible children who were resident on 25 June 2013 and in children for whom we obtained a TST result. The p values are Wald p values from a logistic regression model with a single categorical exposure and having a TST result as the outcome. The model did not account for clustering by school or household.

*Table 5. The distribution of putative risk factors among children for whom I had a TST result and those for whom I did not.*

| | Number of children with TST result (column %) | Number of children without a TST result (column %) | p value |
|---|---|---|---|
| Overall | 1240 (100.0) | 2272 (100.0) | |
| Attendances at commercial spaces in past week | | | |
| None | 909 (78.4) | 1409 (84.0) | 0.0002 |
| 1-2 | 127 (11.0) | 115 (6.9) | |
| 3+ | 123 (10.6) | 154 (9.1) | |
| Attendances at social spaces in past week | | | |
| None | 1126 (94.8) | 1656 (96.5) | 0.03 |
| 1+ | 62 (5.2) | 61 (3.6) | |
| Trips by public transport in past week | | | |
| None | 936 (78.5) | 1353 (79.1) | 0.92 |
| 1-9 | 143 (12.0) | 203 (11.9) | |
| 10+ | 113 (9.5) | 155 (9.1) | |
| Attendances at church in past week | | | |
| None | 758 (63.6) | 1206 (70.7) | 0.0001 |
| 1+ | 433 (36.4) | 500 (29.3) | |
| Attendances at clinic/hospital in past month | | | |
| None | 1026 (86.5) | 1568 (93.0) | <0.0001 |
| 1+ | 160 (13.5) | 119 (7.1) | |
| Visits to other households in past week | | | |
| None | 729 (64.2) | 1173 (71.4) | 0.0001 |
| 1 | 173 (15.2) | 222 (13.5) | |
| 2+ | 234 (20.6) | 249 (15.2) | |
| Visitors to household in past week | | | |
| None | 803 (70.2) | 1283 (78.0) | <0.0001 |
| 1 | 105 (9.2) | 119 (7.2) | |
| 2+ | 236 (20.6) | 244 (14.8) | |
| Homestead size (number of residents) | | | |
| 1-9 | 438 (35.3) | 765 (33.7) | 0.59 |
| 10-15 | 421 (34.0) | 781 (34.4) | |
| 16+ | 381 (30.7) | 726 (32.0) | |
| Travel time to nearest PHC in minutes (quintiles) | | | |
| Mean 5 | 229 (18.5) | 471 (20.8) | 0.0001 |
| Mean 32 | 212 (17.2) | 488 (21.6) | |
| Mean 55 | 238 (19.3) | 462 (20.4) | |
| Mean 76 | 285 (23.1) | 415 (18.3) | |
| Mean 102 | 272 (22.0) | 427 (18.9) | |
| Household wealth score (quintiles) | | | |
| 1 (poorest) | 260 (21.1) | 339 (15.4) | <0.0001 |
| 2 | 293 (23.8) | 475 (21.5) | |
| 3 | 258 (20.9) | 464 (21.0) | |
| 4 | 242 (19.6) | 487 (22.1) | |
| 5 (richest) | 179 (14.5) | 443 (20.1) | |

| | | | |
|---|---|---|---|
| 2009 area level deprivation in 3km Gaussian kernel around homestead | | | |
| 1 (poorest) | 311 (25.7) | 375 (17.0) | <0.0001 |
| 2 | 269 (22.2) | 413 (18.7) | |
| 3 | 229 (18.9) | 455 (20.6) | |
| 4 | 211 (17.4) | 473 (21.4) | |
| 5 (richest) | 190 (15.7) | 493 (22.3) | |
| 2009 number of HIV positive individuals in homestead | | | |
| 0 | 504 (41.7) | 981 (44.4) | 0.15 |
| 1 | 423 (35.0) | 703 (31.8) | |
| 2+ | 283 (23.4) | 525 (23.8) | |
| 2009 HIV prevalence in 3km Gaussian kernel around homestead (tertiles) | | | |
| Mean 15% | 468 (38.7) | 672 (30.4) | <0.0001 |
| Mean 21% | 424 (35.0) | 717 (32.5) | |
| Mean 29% | 318 (26.3) | 820 (37.1) | |
| Population density in 3km Gaussian kernel around homestead (quartiles) | | | |
| Mean 60 residents/km$^2$ | 390 (32.2) | 467 (21.1) | <0.0001 |
| Mean 126 residents/km$^2$ | 353 (29.2) | 500 (22.6) | |
| Mean 290 residents/km$^2$ | 254 (21.0) | 601 (27.2) | |
| Mean 975 residents/km$^2$ | 213 (17.6) | 641 (29.0) | |

Whilst some of the p values were low, there was little real difference in the distribution of risk factors between children for whom we had a TST result and other resident eligible children. I was slightly more likely to recruit children resident in areas with a lower population density and from more deprived communities and households.

### Minimally adjusted estimates

All data presented in this section (Table 6) are from mixed effect logistic regression models accounting with random intercepts for clustering by homestead (the one in which children lived for the longest period of time) and by school. They each contain a single categorical exposure variable and no covariates.

*Table 6. Minimally adjusted associations between putative risk factors and MTB infection.*

| | Infection defined as a non zero TST | | | Infection defined as TST ≥10mm | | |
|---|---|---|---|---|---|---|
| | n/N (%) | OR (95% CI) [1] | p value [2] | n/N (%) | OR (95% CI) [1] | p value [2] |
| Age at testing in years | | | | | | |
| 6 | 26/144 (18) | Referent | 0.93 | 17/144 (12) | Referent | 0.25 |
| 7 | 113/676 (17) | 0.96 (0.52-1.76) | | 85/676 (13) | 1.33 (0.60-2.13) | |
| 8 | 75/420 (18) | 1.04 (0.55-1.97) | | 52/420 (12) | 1.11 (0.57-2.15) | |
| Sex | | | | | | |
| Male | 98/622 (16) | Referent | 0.18 | 71/622 (11) | Referent | 0.28 |
| Female | 116/618 (19) | 1.30 (0.88-1.91) | | 83/618 (13) | 1.24 (0.84-1.82) | |
| Attendances at commercial spaces in past week | | | | | | |
| None | 153/909 (17) | Referent | 0.95 | 115/909 (13) | Referent | 0.43 |
| 1-2 | 21/127 (17) | 0.98 (0.53-1.82) | | 11/127 (9) | 0.64 (0.32-1.29) | |
| 3+ | 23/123 (19) | 1.11 (0.60-2.05) | | 15/123 (12) | 0.93 (0.49-1.75) | |
| Attendances at social spaces in past week | | | | | | |
| None | 191/1126 (17) | Referent | 0.60 | 136/1126 (12) | Referent | 0.57 |
| 1+ | 12/62 (19) | 1.26 (0.53-2.99) | | 9/62 (15) | 1.28 (0.55-2.97) | |
| Trips by public transport in past week | | | | | | |
| None | 168/936 (18) | Referent | 0.28 | 122/936 (13) | Referent | 0.25 |
| 0-9 | 24/143 (17) | 0.95 (0.52-1.76) | | 16/143 (11) | 0.87 (0.46-1.62) | |
| 10+ | 12/113 (11) | 0.54 (0.24-1.21) | | 8/113 (7) | 0.52 (0.22-1.19) | |
| Attendances at church in past week | | | | | | |
| None | 131/758 (17) | Referent | 0.86 | 96/758 (13) | Referent | 0.53 |
| 1+ | 74/433 (17) | 0.96 (0.64-1.46) | | 50/433 (12) | 0.88 (0.57-1.33) | |
| Attendances at clinic/hospital in past month | | | | | | |
| None | 174/1026 (17) | Referent | 0.75 | 121/1026 (12) | Referent | 0.28 |
| 1+ | 29/160 (18) | 1.09 (0.63-1.91) | | 24/160 (15) | 1.36 (0.78-2.35) | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Visits to other households in past week | | | | | | |
| None | 122/729 (17) | Referent | 0.05 | 87/729 (12) | Referent | 0.07 |
| 1 | 24/173 (14) | 0.72 (0.35-1.51) | | 15/173 (9) | 0.62 (0.28-1.40) | |
| 2+ | 53/234 (23) | 1.80 (0.95-3.41) | | 38/234 (16) | 1.61 (0.85-3.04) | |
| Visitors to household in past week | | | | | | |
| None | 134/803 (17) | Referent | 0.27 | 95/803 (12) | Referent | 0.26 |
| 1 | 14/105 (13) | 0.74 (0.30-1.83) | | 9/105 (9) | 0.69 (0.28-1.69) | |
| 2+ | 48/236 (20) | 1.49 (0.81-2.75) | | 35/236 (15) | 1.42 (0.81-2.50) | |
| Homestead size (number of residents) | | | | | | |
| 1-9 | 88/438 (20) | Referent | 0.26 | 56/438 (13) | Referent | 0.70 |
| 10-15 | 67/421 (16) | 0.72 (0.46-1.14) | | 48/421 (11) | 0.89 (0.55-1.43) | |
| 16+ | 59/381 (15) | 0.72 (0.45-1.14) | | 50/381 (13) | 1.10 (0.68-1.79) | |
| Travel time to nearest PHC in minutes (quintiles) | | | | | | |
| Mean 5 | 47/229 (21) | Referent | 0.46 | 38/229 (17) | Referent | 0.09 |
| Mean 32 | 42/212 (20) | 0.98 (0.52-1.85) | | 33/212 (16) | 0.94 (0.52-1.71) | |
| Mean 55 | 39/238 (16) | 0.72 (0.38-1.39) | | 28/238 (12) | 0.65 (0.35-1.22) | |
| Mean 76 | 45/285 (16) | 0.67 (0.35-1.28) | | 29/285 (10) | 0.54 (0.28-1.02) | |
| Mean 102 | 41/272 (15) | 0.61 (0.31-1.19) | | 26/272 (10) | 0.49 (0.25-0.94) | |
| Household wealth score (quintiles) | | | | | | |
| 1 (poorest) | 45/260 (17) | Referent | 0.19 | 28/260 (11) | Referent | 0.19 |
| 2 | 44/293 (15) | 0.80 (0.44-1.46) | | 32/293 (11) | 1.00 (0.55-1.83) | |
| 3 | 57/258 (22) | 1.55 (0.85-2.85) | | 44/258 (17) | 1.84 (1.00-3.37) | |
| 4 | 36/242 (15) | 0.87 (0.46-1.63) | | 28/242 (12) | 1.14 (0.60-2.14) | |
| 5 (richest) | 31/179 (17) | 1.14 (0.58-2.27) | | 22/179 (12) | 1.26 (0.63-2.50) | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2009 area level deprivation in 3km Gaussian kernel around homestead | | | | | | |
| 1 (poorest) | 47/311 (15) | Referent | 0.25 | 32/311 (10) | Referent | 0.34 |
| 2 | 47/269 (17) | 1.30 (0.70-2.41) | | 31/269 (12) | 1.24 (0.66-2.31) | |
| 3 | 41/299 (18) | 1.48 (0.75-2.90) | | 29/229 (13) | 1.47 (0.75-2.86) | |
| 4 | 45/211 (21) | 1.94 (0.96-3.90) | | 34/211 (16) | 1.96 (1.00-3.85) | |
| 5 (richest) | 27/109 (14) | 0.94 (0.43-2.06) | | 23/190 (12) | 1.26 (0.60-2.68) | |
| 2009 number of HIV positive individuals in homestead | | | | | | |
| 0 | 85/504 (17) | Referent | 0.95 | 53/504 (11) | Referent | 0.06 |
| 1 | 71/423 (17) | 1.03 (0.65-1.61) | | 49/423 (12) | 1.13 (0.71-1.78) | |
| 2+ | 51/283 (18) | 1.09 (0.66-1.80) | | 47/283 (17) | 1.75 (1.07-2.88) | |
| 2009 HIV prevalence in 3km Gaussian kernel around homestead (tertiles) | | | | | | |
| Mean 15% | 76/468 (16) | Referent | 0.62 | 53/468 (11) | Referent | 0.68 |
| Mean 21% | 79/424 (19) | 1.25 (0.78-2.01) | | 53/424 (13) | 1.13 (0.71-1.82) | |
| Mean 29% | 52/318 (16) | 1.06 (0.61-1.83) | | 43/318 (14) | 1.26 (0.74-2.17) | |
| Population density (quartiles) | | | | | | |
| Mean 60 residents/km$^2$ | 71/390 (18) | Referent | 0.98 | 50/390 (13) | Referent | 0.98 |
| Mean 126 residents/km$^2$ | 61/353 (17) | 0.96 (0.57-1.62) | | 40/353 (11) | 0.94 (0.55-1.62) | |
| Mean 290 residents/km$^2$ | 42/254 (17) | 0.88 (0.48-1.62) | | 32/254 (13) | 1.03 (0.57-1.86) | |
| Mean 975 residents/km$^2$ | 33/213 (15) | 0.89 (0.44-1.80) | | 27/213 (13) | 1.08 (0.56-2.12) | |
| History of TB contact in the household | | | | | | |
| No | 145/941 (15) | Referent | 0.002 | 104/941 (11) | Referent | 0.007 |
| Yes | 67/280 (24) | 2.04 (1.24-3.38) | | 49/280 (18) | 1.82 (1.15-2.86) | |

*1. From a mixed effects logistic regression model accounting for clustering by household and school*
*2. Likelihood ratio test.*

### *Age and sex*

The expected trend towards increased odds of MTB infection with age was not seen, presumably as there was limited variation in age in the sample. There was no clear difference in infection prevalence by sex.

### *Social contact*

There was little evidence in the unadjusted analysis that these social contact variables were associated with MTB infection using either threshold. The exception was social contact within households with weak evidence that visits to other households were associated with increased odds of MTB infection (p=0.05 with the lower threshold and p=0.07 with the higher threshold).

I planned *a priori* to create aggregate measures of the number of contacts in household spaces, non clinical public spaces and clinical public spaces, assuming that effect estimates of the variables I planned to combine were not in opposite directions in unadjusted analyses. This was, broadly speaking, the case. However, given a trend towards reduced risk of MTB infection with greater use of public transport (not statistically significant), I decided not to include public transport in the aggregate measure of contacts in non-clinical public spaces.

### *Socioeconomic position and access to healthcare*

Here, there was a trend towards reduced odds of MTB infection in children residing further from clinics. This trend was the opposite to that I was expecting. This might be a result of confounding by, e.g., other aspects of the more urban environments in which clinics tend to be located. Using the higher threshold for TST positivity (but not

the lower threshold), there was weak statistical evidence that this was not a chance finding (p=0.09).

There was no evidence of an association between either household or community level socioeconomic position and MTB infection in children. There was no evidence for an interaction between household wealth scores and travel time to PHC (p=0.94 using all positive results and 0.18 using a ≥10mm threshold).

### *HIV prevalence*

To obtain the number of HIV positive adults in the homestead in which the child lived longest, I took the result of the HIV test for each adult obtained closest to 2009, as described above. However, by this method 44% of adults registered in the surveillance programme still had no HIV test result and 13% of households contained no adult with an HIV test result.

There was weak evidence using the higher TST threshold that exposure to HIV positive adults in the household might be associated with MTB infection in the child (p=0.06) but no suggestion of such an association using the lower TST threshold. There was little evidence that HIV prevalence in the community immediately surrounding the household was associated with MTB infection in children.

### *Population density*

There was no evidence in this model for an association between population density and odds of MTB infection.

### *TB contact*

Reassuringly, there was strong evidence for increased odds of MTB infection in children with a history of household TB contact.

## Adjusted estimates

### *Socioeconomic position and access to healthcare*

*Table 7. Adjusted associations of distance to nearest clinic and measures of socioeconomic position with MTB infection (model not adjusted for population density).*

| | Infection defined as a non-zero TST | | Infection defined as TST ≥10mm | |
|---|---|---|---|---|
| | aOR (95% CI) | p value [1] | aOR (95% CI) | p value [1] |
| Travel time to nearest PHC in minutes (quintiles) | | | | |
| Mean 5 | Referent | 0.72 | Referent | 0.26 |
| Mean 32 | 0.88 (0.44-1.75) | | 0.86 (0.44-1.67) | |
| Mean 55 | 0.68 (0.32-1.43) | | 0.61 (0.29-1.28) | |
| Mean 76 | 0.62 (0.28-1.36) | | 0.48 (0.21-1.09) | |
| Mean 102 | 0.59 (0.25-1.42) | | 0.43 (0.17-1.05) | |
| Household wealth score (quintiles) | | | | |
| 1 (poorest) | Referent | 0.39 | Referent | 0.51 |
| 2 | 0.73 (0.39-1.36) | | 0.93 (0.49-1.75) | |
| 3 | 1.22 (0.65-2.32) | | 1.48 (0.77-2.86) | |
| 4 | 0.73 (0.37-1.46) | | 0.95 (0.47-1.93) | |
| 5 (richest) | 0.94 (0.45-1.97) | | 0.99 (0.46-2.13) | |
| 2009 area level deprivation in 3km Gaussian kernel around homestead | | | | |
| 1 (poorest) | Referent | 0.36 | Referent | 0.71 |
| 2 | 1.21 (0.62-2.36) | | 1.02 (0.51-2.02) | |
| 3 | 1.21 (0.57-2.56) | | 0.99 (0.47-2.11) | |
| 4 | 1.41 (0.61-3.24) | | 1.10 (0.48-2.52) | |
| 5 (richest) | 0.66 (0.25-1.76) | | 0.65 (0.25-1.70) | |

*1. Likelihood ratio test.*

*Table 8. Adjusted associations of distance to nearest clinic and measures of socioeconomic position with MTB infection (model adjusted for population density).*

| | Infection defined as a non-zero TST | | Infection defined as TST ≥10mm | |
|---|---|---|---|---|
| | aOR (95% CI) | p value [1] | aOR (95% CI) | p value [1] |
| Travel time to nearest PHC in minutes (quintiles) | | | | |
| Mean 5 | Referent | 0.77 | Referent | 0.29 |
| Mean 32 | 0.89 (0.44-1.82) | | 0.87 (0.44-1.72) | |
| Mean 55 | 0.68 (0.32-1.47) | | 0.62 (0.29-1.31) | |
| Mean 76 | 0.65 (0.29-1.47) | | 0.49 (0.21-1.14) | |
| Mean 102 | 0.61 (0.25-1.46) | | 0.44 (0.18-1.07) | |
| Household wealth score (quintiles) | | | | |
| 1 (poorest) | Referent | 0.40 | Referent | 0.53 |
| 2 | 0.73 (0.39-1.36) | | 0.93 (0.49-1.74) | |
| 3 | 1.19 (0.63-2.26) | | 1.46 (0.76-2.82) | |
| 4 | 0.71 (0.36-1.43) | | 0.94 (0.47-1.91) | |
| 5 (richest) | 0.94 (0.45-1.97) | | 0.99 (0.46-2.13) | |
| 2009 area level deprivation in 3km Gaussian kernel around homestead | | | | |
| 1 (poorest) | Referent | 0.31 | Referent | 0.81 |
| 2 | 1.32 (0.65-2.69) | | 1.07 (0.52-2.21) | |
| 3 | 1.48 (0.64-3.44) | | 1.09 (0.47-2.50) | |
| 4 | 1.62 (0.63-4.15) | | 1.16 (0.47-2.91) | |
| 5 (richest) | 0.58 (0.15-2.34) | | 0.62 (0.16-2.38) | |

*1. Likelihood ratio test.*

There was little evidence that household wealth or area level deprivation were associated with MTB infection in the child in a model also adjusted for age, sex and travel time to the nearest clinic. The addition of population density to the model changed little.

From the same model, there was a trend towards reduced odds of MTB infection with increasing distance to the nearest clinic. This trend was not attenuated by adding population density to the model. However, there was no evidence to support the idea that this trend was anything other than a chance observation (p=0.77 using all positive reactions and p=0.29 using the ≥10mm threshold).

### *Adult HIV prevalence*

*Table 9. Adjusted associations between household and community level HIV prevalence and MTB infection (model not adjusted for population density).*

| | Infection defined as a non-zero TST | | Infection defined as TST ≥10mm | |
|---|---|---|---|---|
| | aOR (95% CI) | p value [1] | aOR (95% CI) | p value [1] |
| 2009 number of HIV positive individuals in homestead | | | | |
| 0 | Referent | 0.90 | Referent | 0.05 |
| 1 | 1.03 (0.66-1.61) | | 1.14 (0.71-1.83) | |
| 2+ | 1.12 (0.68-1.86) | | 1.81 (1.08-3.07) | |
| 2009 HIV prevalence in 3km Gaussian kernel around homestead (tertiles) | | | | |
| Mean 15% | Referent | 0.40 | Referent | 0.76 |
| Mean 21% | 1.24 (0.76-2.00) | | 1.11 (0.68-1.81) | |
| Mean 29% | 0.85 (0.45-1.63) | | 0.90 (0.47-1.71) | |

*1. Likelihood ratio test.*

*Table 10. Adjusted associations between household and community level HIV prevalence and MTB infection (model adjusted for population density).*

| | Infection defined as a non-zero TST | | Infection defined as TST ≥10mm | |
|---|---|---|---|---|
| | aOR (95% CI) | p value [1] | aOR (95% CI) | p value [1] |
| 2009 number of HIV positive individuals in homestead | | | | |
| 0 | Referent | 0.89 | Referent | 0.05 |
| 1 | 1.03 (0.66-1.62) | | 1.14 (0.72-1.83) | |
| 2-7 | 1.13 (0.68-1.87) | | 1.81 (1.08-3.05) | |
| 2009 HIV prevalence in 3km Gaussian kernel around homestead (tertiles) | | | | |
| Mean 15% | Referent | 0.18 | Referent | 0.62 |
| Mean 21% | 1.38 (0.83-2.29) | | 1.17 (0.70-1.95) | |
| Mean 29% | 0.81 (0.41-1.57) | | 0.88 (0.46-1.68) | |

*1. Likelihood ratio test.*

There was weak evidence that exposure to HIV positive adults in the home was associated with increased odds of MTB infection in the child but only using the ≥10mm threshold. This was from a model adjusted for age, sex, household wealth score, area level deprivation and travel time to nearest clinic, and area level HIV prevalence. From the same model, there was no evidence that HIV prevalence in the 3km around the homestead was associated with MTB infection in the child.

I repeated these analyses using alternative thresholds (≥15mm, all positive cases plus reported receipt of TB treatment, and TST reactions of ≥10mm plus reported receipt of TB treatment). I did not adjust for population density as it was not associated with MTB infection and did not appear to be an important confounder in the main analysis.

Using the ≥15mm threshold, the association between exposure to HIV positive adults in the household and MTB infection in the child appeared stronger with children exposed to two or more HIV positive adults in the homestead having 3.50 times the adjusted odds of MTB infection (95% CI 1.33-9.18). However, model fit seemed poor with implausible values predicted for other covariates. Adding children reported to be receiving TB treatment as cases did not alter estimates in either the model including all positive reactions (in which there remained no association) or in the model including reactions of ≥10mm.

Adding history of household TB contact to the model only slightly attenuated the association between exposure to HIV positive adults in the home and MTB infection in the child with children exposed to two or more HIV positive adults having 1.63 times the adjusted odds of MTB infection using the higher TST threshold (95% CI 0.97-2.76).

### *Social contact patterns*

During data cleaning, I noted that there were 65 children where the number of journeys of more than an hour in duration was reported to be greater than the total number of journeys. Nearly all of these nonsense values were captured by four fieldworkers. They had mostly recorded exactly twice as many journeys of more than one hour's duration as the total number of journeys. I spoke to three of the fieldworkers. They had understood the second question to mean total number of journeys including return journeys.

I resolved this issue in the following way

1. Where the total number of journeys was exactly twice the number of journeys more than an hour, I set the total number of journeys to the larger figure and set the number of journeys more than an hour to missing.
2. Where it was not exactly twice the number and where there was no data entry problem, I set both values to missing.
3. I elected not to analyse the data on journeys of more than one hour in duration, as I was not sure how well it was understood.

A regression model containing all variables I considered *a priori* potential confounders of the association between social contact patterns and MTB infection did not converge. I therefore elected to drop the community level deprivation variable and the community level HIV prevalence variable out of the model. Given these were not associated with MTB infection in the adjust analyses presented above, I considered them unlikely to be important confounding variables here.

I also elected not to adjust each social contact variable for the other social contact variables, so as not to over parameterise the model. The associations between the social contact variables were limited although clinic attendees had slightly more community and household contact than children that had not visited clinic or hospital in the preceding month. Note, adjusting clinic attendance for community or household contact did not meaningfully alter the effect estimate for either measure of social contact (data not shown).

*Table 11. Adjusted associations between social contact patterns and MTB infection (model not adjusted for population density).*

| | Infection defined as a non-zero TST | | Infection defined as TST ≥10mm | |
|---|---|---|---|---|
| | aOR (95% CI) | p value [1] | aOR (95% CI) | p value [1] |
| Attendances at non clinical congregate settings in past week | | | | |
| None | Referent | 0.95 | Referent | 0.85 |
| 1 | 0.93 (0.60-1.43) | | 0.91 (0.58-1.43) | |
| 2+ | 0.97 (0.57-1.66) | | 0.87 (0.49-1.52) | |
| Trips by public transport in past week | | | | |
| None | Referent | 0.11 | Referent | 0.14 |
| 0-9 | 0.87 (0.47-1.60) | | 0.84 (0.44-1.60) | |
| 10+ | 0.43 (0.18-1.03) | | 0.44 (0.17-1.09) | |
| Attendances at clinic/hospital in past month | | | | |
| None | Referent | 0.73 | Referent | 0.26 |
| 1+ | 1.10 | | 1.37 (0.79-2.38) | |
| Sum of visits to other households plus number of visitors to own household in past week | | | | |
| None | Referent | 0.39 | Referent | 0.13 |
| 1-2 | 0.97 (0.54-1.73) | | 0.77 (0.42-1.41) | |
| 3+ | 1.51 (0.79-2.89) | | 1.57 (0.85-2.91) | |
| Homestead size (number of residents) | | | | |
| 1-9 | Referent | 0.31 | Referent | 0.72 |
| 10-15 | 0.74 (0.47-1.17) | | 0.91 (0.56-1.49) | |
| 16+ | 0.71 (0.43-1.17) | | 1.11 (0.66-1.87) | |

*1. Likelihood ratio test.*

*Table 12. Adjusted associations between social contact patterns and MTB infection (model adjusted for population density).*

| | Infection defined as a non-zero TST | | Infection defined as TST ≥10mm | |
|---|---|---|---|---|
| | aOR (95% CI) | p value [1] | aOR (95% CI) | p value [1] |
| Attendances at non clinical congregate settings in past week | | | | |
| None | Referent | 0.97 | Referent | 0.89 |
| 1 | 0.95 (0.61-1.48) | | 0.94 (0.59-1.49) | |
| 2+ | 1.00 (0.59-1.70) | | 0.87 (0.49-1.54) | |
| Trips by public transport in past week | | | | |
| None | Referent | 0.21 | Referent | 0.15 |
| 0-9 | 0.88 (0.47-1.63) | | 0.85 (0.44-1.62) | |
| 10+ | 0.50 (0.21-1.15) | | 0.44 (0.17-1.10) | |
| Attendances at clinic/hospital in past month | | | | |
| None | Referent | 0.63 | Referent | 0.24 |
| 1+ | 1.15 (0.66-1.99) | | 1.39 (0.80-2.42) | |
| Sum of visits to other households plus number of visitors to own household in past week | | | | |
| None | Referent | 0.34 | Referent | 0.12 |
| 1-2 | 0.95 (0.53-1.72) | | 0.75 (0.40-1.39) | |
| 3+ | 1.55 (0.80-3.00) | | 1.56 (0.84-2.91) | |
| Homestead size (number of residents) | | | | |
| 1-9 | Referent | 0.34 | Referent | 0.72 |
| 9-15 | 0.74 (0.47-1.18) | | 0.91 (0.55-1.49) | |
| 16+ | 0.72 (0.44-1.18) | | 1.11 (0.66-1.88) | |

*1. Likelihood ratio test.*

In these data, there was very weak evidence (p=0.11 at the lower threshold and p=0.14 at the higher threshold) that children with greater exposure to public transport had lower adjusted odds of MTB infection than children with no exposure to public transport. However, this effect was attenuated and no longer close to statistical significance after adjusting for population density.

In these data, there was very weak evidence at the higher TST threshold (p=0.12 in the model also adjusted for population density) that children who visited other households or had visitors to their household had higher adjusted odds of MTB infection.

## Discussion

### Main findings

In my main analyses locating children to the household in which they lived longest there was weak evidence (p=0.05) that children living in households with more HIV positive individuals were at greater risk of MTB infection. Children living in households with two or more HIV positive adults had nearly twice (aOR 1.81; 95% CI 1.08-3.07) the adjusted odds of MTB infection as children living in households where no adults were known to be HIV positive. This effect persisted after adjusting for population density but was only apparent using the higher TST threshold. At the higher TST threshold, the association was robust to including children reported to be on TB treatment as cases. The association was only partially attenuated by including report of household TB contact in the model.

There was some suggestion (p=0.11 using the lower TST threshold or p=0.14 using the higher threshold) that children with greater exposure to public transport had reduced risk of MTB infection. This effect was most apparent in children reported to have undertaken ten or more journeys by bus or minibus taxi in the preceding week. The majority of children in this group had undertaken exactly ten journeys. Reasons these children might be at reduced risk of MTB infection are discussed below. It should also be noted that this association was attenuated by including population density in the model.

There was a suggestion that children with greater exposure to other households or to members of other households within their home had higher adjusted odds of MTB infection (p=0.13 using the higher TST threshold). However, there was little evidence for this association at the lower TST threshold.

There was also a trend towards reduced odds of MTB infection in children living further from healthcare facilities although there was insufficient evidence to be confident that this was not due to chance.

### Results in context

As outlined above, there is little published literature on the association between wealth and MTB infection and findings have been mixed. That I found no clear association between household wealth and MTB infection at either the household or the community level is interesting given an association between wealth and reduced risk of TB disease is observed consistently in prevalence surveys. As outlined in the introduction, it is plausible that the reduced risk of MTB infection associated with living among adults with lower TB prevalence is offset by living in better constructed (and thus less well ventilated) homes. Alternatively, wealth might facilitate attendance at congregate settings in which there is a risk of MTB transmission.

The association between exposure to a greater number of HIV positive adults in the household and MTB infection in children should be viewed with caution given it was only apparent at the higher TST threshold. However, it is perhaps unsurprising given 76% of TB in the community is among HIV positive individuals(172). This suggests that the data from the South African gold mines in the pre-ART era that suggested HIV positive people with TB made a very limited contribution to MTB transmission(222) may not hold true in this rural community where a growing proportion of HIV positive individuals are on ART(94,240). Studies from other settings exploring this association would be valuable.

That the association between exposure to HIV positive adults in the household and MTB infection in the child was only partly attenuated by adding history of household TB contact into the model might be because report of TB contact was incomplete or, alternatively, that there is some other mechanism by which exposure to HIV positive adults is associated with MTB infection in the child. For example, children might be exposed to TB when accompanying their parents to clinic appointments.

There was a trend towards reduced risk of MTB infection in children living further from healthcare facilities. This was not attenuated by adjusting for population density. This was the opposite of what I expected. One possible explanation would be that healthcare facilities are important sites of MTB transmission. A large proportion of the adult population in the community attend clinics regularly for pre-ART and ART care(94,240). Undiagnosed TB is known to be common in South African healthcare facilities(84,85,87), adherence to TB infection control practices in some South African healthcare facilities is poor(241) and transmission within healthcare facilities was a prominent feature of the Tugela Ferry Extensively Drug Resistant TB (XDR-TB) outbreak(86). Very high force of infection has been described among South African healthcare workers(242). However, in my data, there was no real evidence that this trend towards a reduced risk of MTB infection among children living further from healthcare facilities was anything other than a chance association. Larger studies and studies from other settings are needed.

The trend towards reduced risk of MTB infection with greater exposure to public transportation was also surprising. Wells Riley based modelling undertaken in the Western Cape has predicted that public transport would be an important site of MTB transmission(60,243). It is worth noting that the group of individuals with ten or more

journeys per week were dominated by children who had undertaken exactly ten journeys. These ten journeys were probably journeys to school and back. School runs tend to use 'staff transport' – i.e. the same taxi or truck containing the same individuals (mostly children) each day. As such, these journeys would carry a much lower risk of MTB transmission than if these children hailed a minibus taxi at the roadside where there would be more adults present and children would be exposed to a different set of individuals each day. Children attending school by public transport may be different to their peers in other ways that were not accounted for by the covariates in the model – for example, parents prioritising sending their children to more distant but better schools might have taken other decisions that reduced their children's risk of MTB infection (e.g. better health seeking behaviour).

The trend toward increased odds of MTB infection in children with greater exposure to other households or who lived in households that had a larger number of visitors may be important. The finding in Chapter 2 that a low proportion of TB disease is a result of recent transmission between members of the same household does not necessarily mean that transmission in public spaces is important. An alternative explanation is that people spend a lot of time in other people's homes – private space being used as public space – and that this is an important site of MTB transmission. Wells-Riley based modelling from a Cape Town township suggested household (including other people's households) might be an important site of transmission from adults to children (60). Additional empirical data exploring this association would be valuable.

The lack of an association between other measures of social contact and MTB infection is at odds with what I expected, with the predictions of Wells-Riley based

models(60,243) and with empirical research from other settings(78–80,220). It is possible that public spaces in this community are better ventilated than public spaces elsewhere – the climate is good and windows often kept open. That many of these children reported no attendance at public spaces (other than school, which was a universal exposure for these children) and the potential short duration of exposure to many of these spaces (e.g. shop visits) relative to time spent at home, school and in other people's homes may also explain the observation. Alternative explanations are that relevant social contacts of this sort are difficult to measure, particularly in a cross sectional survey, and that the study was underpowered. Note, that there was a trend towards increased odds of MTB infection in children reporting clinic attendance and that confidence intervals for several other putative risk factors were consistent with both meaningful reductions and increases in the odds of infection. I discuss these issues below.

That neither of my area level variables were associated with MTB infection in the child is notable. This may relate to mobility, as I discuss below. However, I note that a Brazilian study exploring socioeconomic risk factors for TB disease found that household level risk factors predicted risk better than community level risk factors(244).

## Limitations

### *Power*

Due, largely, to failure to make contact with parents and guardians, we obtained TST results on 1240 rather than the intended 1350 resident children. I had insufficient resources to fund repeat visits to the homestead and sending letters home from

school had a low yield. Furthermore, the prevalence of positive tests was lower than observed in the Western Cape, particularly using higher thresholds.

Therefore, the study was underpowered, meaning that I may have missed some associations. In situations where power is limited, there is also a greater chance that observed associations are spurious.

Low power presented a problem in the adjusted analyses, where I was attempting to estimate a large number of parameters relative to the number of positive outcomes observed. There were no obvious problems with model fitting (e.g. large standard errors or strata with no observations). However, a simulation study suggests that biased estimates can be obtained in logistic regression where there are fewer than 10 events per variable, even where models converge without difficulty(245). My adjusted models were all close to this threshold, particularly using the higher TST threshold. Where I attempted a sensitivity analysis using a ≥15mm threshold (only 70 events), I obtained implausible results. The simulation study did not consider a multi-level data structure. Furthermore, most variables considered were binary, whereas several of my variables were categorical meaning that up to four parameters were estimated per variable (a few of my models had fewer than ten events per parameter estimated). For these reasons, I cannot be certain that some of my adjusted analyses were not slightly over parameterised, particularly using the higher TST threshold.

#### *Unmeasured and, as yet, unstudied variables*

By siting my survey in the surveillance area, I was able to look for associations between MTB infection and a number of important putative risk factors. However, besides reported TB contact (almost all household contact), I did not have data on

adult TB prevalence in putative sites of transmission. I did not have data on adult-adult or adult-child social contact. These may be important drivers of MTB transmission in Southern Africa with recently published modelling based on detailed social contact pattern data suggesting that most MTB infections may be a result of contact with adult men(71). These results were driven by higher TB prevalence in men and age and sex assortative mixing.

Ventilation in putative sites of transmission would be expected to modify the MTB infection risk associated with exposure. I did measure ventilation in a number of public spaces in the community in collaboration with researchers at the Bartlett School of Architecture(246). However, we only studied a small number of buildings so generalising our findings would be inappropriate. If there was substantial variation in the ventilation of buildings that might obscure associations between exposure to those spaces and MTB infection.

There are some ecological data suggesting that TB in Southern Africa may be driven, in part, by high rates of TB disease in migrant labourers – particularly miners(166). Whilst no data have been collected on past employment in the mines or past imprisonment (both strong risk factors for TB disease) we plan, in future work, to look for an association between rates of adult in-migration from outside the surveillance area (into households and communities) and MTB infection in children.

### *Measurement error*

The social contact pattern data looked as I expected it to. Churchgoers tended to attend once a week. Many children reported 10 journeys by public transport, which would be consistent with children using taxis to get to school. Most variables were

right skewed with a small number of children reporting more frequent attendance at certain venues.

However, that I found little evidence for an association between social contact patterns and MTB infection in children may be due to inherent difficulties in measuring social contact patterns.

These data were collected from any adult found at home. In only 42% of cases was this a parent or guardian. It is possible, particularly where households were large and the respondent not closely involved in the care of the child, that children's attendance at particular venues might not have been known. The children were quite independent – most walked to school unaccompanied – and may have visited places without informing adults in the household. Ordinary events, such as trips to the shops for groceries, might be easily forgotten.

Our data related to social contacts within the previous week or month whereas MTB infections may have occurred many years ago. With more fixed behaviours – e.g. church attendance, or regular clinic attendance for chronic ailments – this may not matter. However, other social contact patterns are likely to vary from week to week and over the lifetime of the child.

I would expect errors in measuring social contact patterns to be mostly non-differential as latent MTB infection should not affect recall or patterns of social contact. As such, this would bias effect estimates towards the null.

Measurement error is also likely in the HIV prevalence variables. Whilst I attempted to reduce missingness by averaging over a number of years, I did not have test results for 44% of adults registered in the surveillance programme. HIV test uptake in population based surveys has been observed to be missing not at random with individuals knowing themselves to be positive less likely to agree to testing(247,248). Relative differences in community level HIV prevalence across the Africa Centre surveillance area have been shown to be robust to choices made regards the handling of missing data(195,240) (although, to my knowledge, missing not at random mechanisms have not been explored). However, missing data may have led me to underestimate exposure to HIV positive adults within the household. Again, this underestimate would probably be independent of TST positivity and thus to lead to bias towards the null.

### *Scale and location*

For the community level variables, we used Gaussian kernels with a 3km radius. However, little is known about the spatial extent of MTB transmission networks and these might well be setting specific. Were most children to be infected within 500m of their home, the kernels I used might have smoothed over meaningful spatial variation in community level exposures. Were many infections to occur more than 10km from children's homes, the relevant variation in community level exposures might not be appreciable using 3km kernels. There is a literature from ecology showing the importance of scale in obtaining reliable estimates(249,250).

In assigning community level exposures, we located children to their homes. However, particularly if much MTB transmission occurs in the community, household may not be the relevant location. Household location may be a reasonable proxy for

the location of community settings in which transmission occurs. For example, most people attend their nearest primary healthcare clinic (169). However, I cannot discount the possibility that null results for community level variables were because homesteads were not a good proxy for the location in which MTB transmission occurred.

### *Mobility*

In chapter 3, there was a suggestion that spatial patterning of MTB infections was more apparent in analyses restricted to children that had lived most of their lives in one homestead. Misattribution of household and community level variables might occur where children have lived in several different places given we cannot know in which location any MTB infection was acquired. Population mobility might also be expected to bias estimates towards the null.

## Conclusions

I have shown limited evidence for an association between household exposure to HIV positive adults and MTB infection in children. I have also shown some evidence for an association between children visiting other households or having visitors to their household and MTB infection. Given the limitations outlined above, these findings should be considered hypothesis generating rather than definitive.

In the next and final chapter, I discuss some potential approaches to obtaining more definitive answers regards risk factors for MTB infection with a focus on locating important sites of MTB transmission.

# 5. Implications for future research and *M. tuberculosis* control

In this short final chapter, I make recommendations for future research and comment on the implications of my findings for TB control.

## Key findings

In Chapter 2, I showed that strain discordance between co-prevalent cases of TB resident in the same household was observed in several countries. Reasons for this discordance are likely to vary. In low burden countries, reactivation TB will explain a higher proportion discordant pairs than in higher burden countries. In high burden countries and in pockets of high transmission within low burden countries, recent transmission from non-resident individuals is likely to be important.

The systematic review suggested that, in some communities and perhaps more generally, recent transmission between members of the same household may explain only a limited proportion of all TB disease. This should maybe not come as a surprise given, usually, only one TB case is found per household. These data were mostly from adults but limited data suggest that between household transmission may contribute to childhood TB in high burden settings.

That TB cases with different genotypes are commonly found in households in a variety of different settings suggests that the higher prevalence of TB disease found among household TB contacts may be driven both by intrahousehold transmission and by the clustering of risk factors for TB disease within households. In Southern

Africa, the clustering of HIV related immunocompromise within households will be important.

In Chapters 3 and 4, I studied MTB infection in primary school children in rural South Africa.

In Chapter 3, I demonstrated that, whilst mixture analysis was unable to confidently ascribe a threshold that best discriminated between MTB infection and potential non-specific reactions, using reasonable thresholds the annual risk of infection in the Africa Centre surveillance area was approximately two percent. Annual risks of MTB infection derived from TST surveys in school children should only be considered proxies for the force of infection in the wider community. These data suggest that the force of infection in this community in Northern KwaZulu-Natal is very high by global standards. The estimate is lower than that observed in communities in the Western Cape(12,185,209–211), comparable to estimates obtained in Zambia(185), and higher than estimates obtained in surveys in Ghana(186), Malawi(197) and Kenya(194). Estimates of trends in ARTI are insensitive to the cut point used(251). My survey will provide a useful baseline allowing a subsequent study to estimate whether the force of infection in this community is rising or falling.

The spatial and risk factor analyses presented in Chapters 3 and 4 were underpowered. In these data, there was a suggestion that the presence of HIV positive adults in the homestead was associated with increased risk of MTB infection. There was also a suggestion that time spent in other people's homes might be associated with MTB infection. There were also a number of surprising null findings and a number of interesting trends that did not reach statistical significance. As

discussed at the end of the previous chapter, population mobility, difficulties in measuring both social contact patterns and household HIV exposure, and the fact that up to eight years may have elapsed between MTB infection and the measurement of putative risk factors mean null results should not come as a surprise. I discuss alternative study designs that might provide more definitive answers below.

## Approaches to locating MTB transmission in communities

To better target TB control interventions, a more detailed understanding of settings in which between household TB transmission occurs will be needed. Below are some thoughts regards how such data might be obtained.

### Accurate measures of social contact patterns

It is likely that low power and measurement error combined to limit my ability to detect associations between use of public space and MTB infection.

Clearly fixed exposures to sites of putative transmission will be simpler to measure – for example, occupation, church attendance, and regular clinic appointments. Such exposures tend to vary little from month to month or from year to year. Such contacts – regular and with longer durations of exposure – may also be the most relevant to MTB transmission.

However, more every day exposures such as visits to a neighbour's house or occasional trips to very crowded indoor spaces, such as banks or post offices, may

be important determinants of MTB transmission at the population level. Measuring such social contacts may be challenging.

For now, questionnaires and social contact pattern diaries are likely to remain the primary means of quantifying exposure to putative sites of transmission. Palwasha Khan took this approach in Karonga, Malawi, and has been able to describe plausible associations between MTB infection in young children and exposure to specific indoor (but not outdoor) congregate settings(220). By studying younger (and therefore less independent) children, she may have obtained more accurate reports of their social contact patterns from adult informants.

There is, however, emerging technology that may allow more accurate estimation of exposure to indoor public spaces. For example, wearable Global Positioning System tags have been used by Robin Wood's group in Cape Town to quantify exposure to specific indoor public spaces and to concurrently estimate ventilation using a $CO_2$ meter (159,252). Other such approaches to measuring social contact patterns, including the use of wearable Radio Frequency ID tags, are being explored(253–255). These technologies are exciting but they would need to be scalable to a few thousand individuals if one wanted to demonstrate an association between social contact patterns, measured in this way, and MTB infection. An alternative approach would be to use these tools to optimise and validate questionnaires or social contact pattern diaries.

### Measuring incident rather than prevalent infection

Another important difference between my study and the results recently reported by Palwasha Khan is that she measured incident MTB infection (with a baseline and

then a follow up TST survey). It is therefore likely that measured social contact patterns will have related more closely to exposure during the period of risk. Clearly, measuring incident infection requires larger studies. However, the observation that the incidence of MTB infections measured longitudinally is higher than that estimated from cross sectional surveys(12,21), might attenuate some of the loss of power. If resources are available, this should be the preferred approach.

An added advantage of infection incidence cohorts is that adolescents and young adults might be included and inference made about the locations in which they acquire MTB infection. A growing body of evidence suggests that adults – especially men – may be the source of most new infections in high burden communities(71). Therefore, understanding where adults acquire MTB infection may be particularly valuable.

## Implications for TB control

### Household case finding

If little TB disease is a result of recent transmission in the household, should we invest resources in household case finding? Fifty years ago, Narain studied the distribution of TB cases in a community in South India and concluded that the approach was misguided.

> *Clinical experience has led to a strong belief that tuberculosis is a family disease and contact examination is a sine qua non for case-finding programmes. Considerable doubts are cast on the usefulness of contact examination in tuberculosis control by the present study, which is based on a sample of an entire population rather than on family contacts of known cases only. Cases of tuberculosis occurred mostly singly in households, and contact examination could have revealed only a very small percentage of the cases in this community.*
>
> Narain et al, 1966(153)

I disagree with Narain that 'considerable doubts are cast on the usefulness of contact examination' by this observation. Even in settings with a very high prevalence of TB disease, the yield of household contact tracing will be higher than active case finding in the general community(133). Even if active disease is not found, contacts may benefit from preventative therapy. In areas where HIV is prevalent, household case finding interventions are also an effective means of finding undiagnosed cases of HIV(256). Household contact tracing should be undertaken where resources allow.

The extent to which household contact tracing impacts on MTB transmission at the community level, however, is an important question. Katharina Kranzer's systematic review found only five studies examining the impact on TB epidemiology of adding active screening to passive case finding(257). Of these, only the ZAMSTAR cluster randomised trial, conducted in 16 communities in Zambia and 8 in the Western Cape of South Africa, allowed a clean comparison of passive case finding with active case finding interventions. The interventions studied included a package of interventions, including TB case finding, HIV testing and the provision of IPT to eligible contacts,

that were delivered in the homes of TB cases(256,258). The outcomes measured included TB prevalence at the end of the trial and incident MTB infection, measured using the TST in a cohort of school children. In the communities randomised to receive the household intervention versus those randomised not to receive the household intervention (the study used a factorial design) the adjusted TB prevalence ratio was 0·82 (0·64–1·04). The adjusted rate ratio for incident TB infection was 0·45 (0·20–1·05)(256). There was thus weak evidence that a package of interventions, including household case finding, had reduced MTB transmission in the intervention communities. Of note, only 4.2% of individuals in the communities allocated to the household intervention arms of the trial were actively screened for tuberculosis.

Modelling also suggests household based interventions can reduce TB incidence at a population level(259). These data were obtained using an agent-based model, parameterised to reflect the epidemiology and household structure seen in Brazil and India. Transmission was partitioned into household and community components. In the model, the lower bound of the proportion of MTB transmission that occurs in households was based on the 19% estimate obtained by Verver in Cape Town(77). In this 'lower bound' scenario, the model suggested that a 10% reduction in incidence could be achieved over five years if 100% of household contacts were screened with a test with 100% sensitivity. Substantial additional benefit was predicted if contacts without TB disease received preventive therapy. Were only 50% of contacts to be screened with a test with 50% sensitivity, the year on year decline in the incidence of TB disease would fall from 2% to 0.5%. Globally TB incidence is estimated to be declining at 1.5% per annum(1).

However, Narain is probably correct that transmission between rather than within households is the primary driver of incident disease in the community. As Rose noted, 'a large number of people at a small risk may give rise to more cases of disease than the small number who are at a high risk' (260). Whilst household contacts are at elevated risk, they comprise a relatively small proportion of the population. Furthermore, in communities where household TB contact is more common, the association between household TB contact and MTB infection is modest – presumably as a result of a high force of infection in the community(154). As such, we should consider whether, particularly in high burden settings, interventions designed to directly interrupt MTB transmission between households might usefully supplement interventions based on passive case finding and household contact tracing.

## Potential approaches to interrupting *M. tuberculosis* transmission in public spaces

### Targeting active case finding interventions

Were specific indoor congregate settings found to be important sites of MTB transmission, transmission might be reduced by targeting active case finding to those settings. The strategy would have two obvious advantages. First, the yield of screening might be higher as individuals in these spaces would have been exposed to a higher force of infection. Second, individuals frequenting such spaces would be expected to contribute more to onward transmission than other individuals. Removing them from the pool of infectious individuals would thus be expected to have a disproportionate impact on MTB transmission.

### TB infection control interventions

TB infection control is a badly neglected area of TB research. There has, to my knowledge, been only one controlled trial of a TB infection control intervention with MTB infection in people as its outcome(261) and none undertaken in high burden countries.

TB infection control interventions comprise administrative controls, e.g. cough triage; environmental controls, such as improvements to ventilation or the installation of upper room ultraviolet germicidal irradiation (UVGI); and the use of personal protective equipment, such as N95 respirators. TB infection control interventions are poorly implemented even in high risk environments such as South African healthcare facilities(241) and generally not considered in non clinical public spaces. Whether such interventions might be implemented in non-clinical public spaces found to be important sites of MTB transmission would depend on the nature of the spaces and their users.

If much MTB transmission occurred in a small number of large institutional spaces, then many of the infection control interventions implemented in healthcare facilities may be feasible. Workers in the space could be regularly screened for TB and HIV positive workers assigned to lower risk activities. Environmental measures, such as improvements in natural ventilation or the installation of UVGI, might be expected to substantially reduce transmission(74,76,262,263). In settings with a very high force of infection, N95 respirators could be considered.

However, if MTB transmission were found to largely occur in small institutions or in gatherings in private spaces then alternative strategies might be needed. For

example, licensing regulations could mandate a minimum level of natural ventilation and custodians of buildings supported to fit extra windows. Alternatively, TB infection control specialists could proactively involve themselves in major construction projects, ensuring new buildings are well ventilated. An example of such a project would be South Africa's Reconstruction and Development Programme which, since 1994, has built more than a million new homes.

Whilst there may be limits to the protection that can be achieved through environmental controls – especially where duration of exposure is long(61) – mathematical models of MTB transmission suggest that TB incidence is very sensitive to changes in the effective contact rate(264). Synergism with case finding interventions might be expected with any missed cases less able to transmit to others.

I have argued elsewhere that such an approach to TB control has a number of inherent advantages(3). Most importantly, it does not place additional demands upon healthcare workers, who are in short supply in many high burden communities.

## Reflections on undertaking research in this setting

Undertaking research at the Africa Centre was brilliant. The area is beautiful. I learnt a huge amount TB, HIV and the practicalities of undertaking population based research. Driving pick up trucks on dirt roads and spending time with six to eight year olds was a lot of fun. My colleagues were great.

The primary challenges in undertaking this research in Northern KwaZulu-Natal were ensuring adequate follow up for children with positive TSTs; managing the

considerable demand for a “TB test” given the study had limited resources; and challenges in making contact with parents and guardians to obtain consent for testing.

I planned for children with positive TSTs and risk factors for TB disease to be reviewed by a doctor at Hlabisa Hospital, the only Government facility in the area able to provide a chest x-ray. The letter in both English and isiZulu, handed to children with positive TSTs, informed parents and guardians that there were free Government buses running to the hospital from local clinics three mornings a week and that we would provide money for the return journey. I left a box of cash in TB clinic for this purpose. I planned to undertake clinical work in the hospital to compensate them for the additional workload my study generated.

In reality, things were more complex. For a period, there were problems with the buses. To address this, we provided families with a phone number to call should they have problems with transport and ended up providing transport in Africa Centre vehicles to a small number of children. Delays in obtaining my medical license meant that I ended up doing my clinical work after my study had concluded.

In marked contrast to HIV research in adults, where the Africa Centre has struggled with high test refusal rates(238), parents and guardians seemed very keen to obtain a “TB test” for their children. My uncertainty regards the number of TST tests my research nurse would be able to undertake in a day led me to restrict recruitment to Grade 1 and 2 students. In reality, my research nurse was excellent and we could have tested substantially more children each day. Families and schools were often keen for us to test older children and, in one school, parents and the headmistress

argued that not providing the test to older children was unjust. There was little that could be done once the study was underway except to explain that most children with positive TSTs would be healthy and that unwell older children could be evaluated for free at local healthcare facilities.

Finally, consent rates were lower than expected with letters sent home from school resulting in the enrolment of a low proportion of children whose parents or guardians had not been found at home by the fieldworkers during surveillance visits. This strategy might have been more successful had we had more success finding children at school. I would, in future, invest more resources in contacting parents or guardians at home. For example, I would fund repeat visits by fieldworkers to the homestead. The Africa Centre should consider capturing data on which school children attend as part of their routine surveillance. This would greatly assist researchers attempting to recruit to school based studies by sending letters home from schools.

## Conclusions

In this thesis I have undertaken a systematic review which provides some evidence to support a growing body of research suggesting that much (perhaps most) TB disease results from transmission between rather than within households. In one rural community in Northern KwaZulu-Natal, I demonstrated a very high force of infection. In that community, I was frustratingly unable to demonstrate associations between exposure to specific public spaces and odds of MTB infection. My inability to do so was probably a result of limited power and limitations of the study design. However, it is, of course, possible that exposure to public spaces in this community is not an important determinant of MTB infection in children, even if it appears to be elsewhere. Achieving a better understanding of where MTB transmission occurs in

high burden settings should be a research priority as it would permit TB control activities to be better targeted. I believe cohorts capturing detailed social contact pattern data and measuring MTB infection longitudinally is one means by which that might be achieved.

## 6. References


1. World Health Organization. Global tuberculosis report 2015. Geneva; 2015.
2. Wood R, Lawn SD, Johnstone-Robertson S, Bekker L. Tuberculosis control has failed in South Africa – time to reappraise strategy. S Afr Med J. 2011;101(2):2009–12.
3. Yates TA, Tanser F, Abubakar I. Plan Beta for Tuberculosis: it's time to think seriously about poorly ventilated congregate settings. Int J Tuberc Lung Dis. 2016;20(1):5–10.
4. Tanser F, Hosegood V, Bärnighausen T, Herbst K, Nyirenda M, Muhwava W, et al. Cohort Profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. Int J Epidemiol. 2008 Oct;37(5):956–62.
5. Dye C, Bassili A, Bierrenbach a L, Broekmans JF, Chadha VK, Glaziou P, et al. Measuring tuberculosis burden, trends, and the impact of control programmes. Lancet Infect Dis. 2008 Apr;8(4):233–43.
6. World Health Organisation. Guidelines on the management of latent tuberculosis infection. Geneva; 2015.
7. Rieder H. Epidemiologic Basis of Tuberculosis Control. 1st. ed. Paris: International Union Against Tuberculosis and Lung Disease; 1999.
8. Edwards L, Acquaviva F, Livesay V, Cross F, Palmer C. An atlas of sensitivity to tuberculin, PPD-B, and histoplasmin in the United States. Am Rev Respir Dis. 1969;99:1–132.
9. Menzies D. What does tuberculin reactivity after bacille Calmette-Guérin vaccination tell us? Clin Infect Dis. 2000;31 Suppl 3:S71–4.
10. Farhat M, Greenaway C, Pai M, Menzies D. False-positive tuberculin skin tests : what is the absolute effect of BCG and non-tuberculous mycobacteria ?

Int J Tuberc Lung Dis. 2006;10(March):1192–204.

11. Menzies D, Pai M, Comstock G. Meta-analysis : New Tests for the Diagnosis of Latent Tuberculosis Infection : Areas of Uncertainty and Recommendations for Research. Ann Intern Med. 2007;146:340–54.

12. Andrews JR, Hatherill M, Mahomed H, Hanekom WA, Campo M, Hawn TR, et al. The dynamics of QuantiFERON®-TB Gold In-Tube conversion and reversion in a cohort of South African adolescents. Am J Respir Crit Care Med. 2015;191(5):584–91.

13. du Plessis DG, Warren R, Richardson M, Joubert JJ, van Helden PD. Demonstration of reinfection and reactivation in HIV-negative autopsied cases of secondary tuberculosis: multilesional genotyping of Mycobacterium tuberculosis utilizing IS 6110 and other repetitive element-based DNA fingerprinting. Tuberculosis (Edinb). 2001;81(3):211–20.

14. Thomas TA, Mondal D, Noor Z, Liu L, Alam M, Haque R, et al. Malnutrition and Helminth Infection Affect Performance of an Interferon γ–Release Assay. Pediatrics. 2010;126(6):e1522–9.

15. Mandalakas AM, van Wyk S, Kirchner HL, Walzl G, Cotton M, Rabie H, et al. Detecting tuberculosis infection in HIV-infected children: a study of diagnostic accuracy, confounding and interaction. Pediatr Infect Dis J. 2013 Mar;32(3):e111–8.

16. Verrall AJ, G. Netea M, Alisjahbana B, Hill PC, van Crevel R. Early clearance of Mycobacterium tuberculosis : a new frontier in prevention. Immunology. 2014 Apr 11;141(4):506–13.

17. Cobat A, Gallant CJ, Simkin L, Black GF, Stanley K, Hughes J, et al. Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. J Exp Med. 2009;206(12):2583–91.

18. Cobat A, Barrera LF, Henao H, Arbeláez P, Abel L, García LF, et al. Tuberculin skin test reactivity is dependent on host genetic background in Colombian tuberculosis household contacts. Clin Infect Dis. 2012;54(7):968–71.

19. Cobat A, Poirier C, Hoal E, Boland-Auge A, de La Rocque F, Corrard F, et al. Tuberculin Skin Test Negativity Is Under Tight Genetic Control of Chromosomal Region 11p14-15 in Settings With Different Tuberculosis Endemicities. J Infect Dis. 2015;211(2):317–21.

20. Jabot-Hanin F, Cobat A, Feinberg J, Grange G, Remus N, Poirier C, et al. Major loci on chromosomes 8q and 3q control BCG and ESAT-6 triggered IFN-γ production, respectively, in various populations. J Infect Dis. 2016;213(7):1173–9.

21. Fine PEM, Bruce J, Ponnighaus JM, Nkhosa P, Harawa A, Vynnycky E. Tuberculin sensitivity : conversions and reversions in a rural African population. Int J Tuberc Lung Dis. 1999;3(May):962–75.

22. Young DB, Gideon HP, Wilkinson RJ. Eliminating latent tuberculosis. Trends Microbiol. 2009 May;17(5):183–8.

23. Barry CE, Boshoff HI, Dartois V, Dick T, Ehrt S, Flynn J, et al. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. Nat Rev Microbiol. 2009 Dec;7(12):845–55.

24. Lin PL, Flynn JL. Understanding latent tuberculosis: a moving target. J Immunol. 2010 Jul 1;185(1):15–22.

25. Esmail H, Barry CE, Young DB, Wilkinson RJ. The ongoing challenge of latent tuberculosis. Philos Trans R Soc Lond B Biol Sci. 2014 Jan;369(1645):20130437.

26. Goo JM, Im JG, Do KH, Yeo JS, Seo JB, Kim HY, et al. Pulmonary

tuberculoma evaluated by means of FDG PET: findings in 10 cases. Radiology. 2000;216(1):117–21.

27. Ayele WY, Neill SD, Zinsstag J, Weiss MG, Pavlik I. Bovine tuberculosis: an old disease but a new threat to Africa. Int J Tuberc Lung Dis. 2004;8(8):924–37.

28. Fennelly KP, Jones-lópez EC. Quantity and quality of inhaled dose predicts immunopathology in tuberculosis. Front Immunol. 2015;6(June):1–13.

29. Koch R. The aetiology of tuberculosis. A translation by Berna Pinner and Max Pinner. Am Rev Tuberc. 1932;25:284–323.

30. Wells WF. On air-borne infection. Study II. Droplets and droplet nuclei. Am J Hyg. 1934;20(3):611–8.

31. Riley RL, Wells WF, Mills CC, Nyka W, McLean RL. Air hygiene in tuberculosis: quantitative studies of infectivity and control in a pilot ward. Am Rev Tuberc. 1957;75(3):420–31.

32. Wells WF, Ratcliffe HL, Crumb C. On the mechanics of droplet nuclei infection. II. Quantitative experimental air-borne tuberculosis in rabbits. Am J Hyg. 1948;47(1):11–28.

33. Lurie MB, Heppleston AC, Abramson S, Swartz IB. An evaluation of the method of quantitative airborne infection and its use in the study of the pathogenesis of tuberculosis. Am Rev Tuberc. 1950;61(6):765–97.

34. Houk VN. Spread of Tuberculosis via recirculated air in a naval vessel: the Byrd study. Ann N Y Acad Sci. 1980;353:10–24.

35. Adetifa IMO, Ota MOC, Jeffries DJ, Hammond A, Lugos MD, Donkor S, et al. Commercial interferon gamma release assays compared to the tuberculin skin test for diagnosis of latent Mycobacterium tuberculosis infection in childhood contacts in the Gambia. Pediatr Infect Dis J. 2010 May;29(5):439–43.

36. Fennelly KP, Jones-López EC, Ayakaka I, Kim S, Menyha H, Kirenga B, et al. Variability of infectious aerosols produced during coughing by patients with pulmonary tuberculosis. Am J Respir Crit Care Med. 2012 Sep 1;186(5):450–7.

37. Jones-López EC, Namugga O, Mumbowa F, Ssebidandi M, Mbabazi O, Moine S, et al. Cough aerosols of Mycobacterium tuberculosis predict new infection: a household contact study. Am J Respir Crit Care Med. 2013 May 1;187(9):1007–15.

38. Turner RD, Bothamley GH. Cough and the transmission of tuberculosis. J Infect Dis. 2014;44(0):1–19.

39. Loudon R, Roberts R. Droplet expulsion from the respiratory tract. Am Rev Respir Dis. 1967;95:435–42.

40. Loudon RG, Spohn SK. Cough frequency and infectivity in patients with pulmonary tuberculosis. Am Rev Respir Dis. 1969;99(1):109–11.

41. Jones-López EC, Kim S, Fregona G, Marques-Rodrigues P, Hadad DJ, Molina LPD, et al. Importance of cough and M. tuberculosis strain type as risks for increased transmission within households. PLoS One. 2014;9(7):e100984.

42. Loudon R, Roberts R. Singing and the Dissemination of Tuberculosis. Am Rev Respir Dis. 1968;98:297–300.

43. Morawska L, Johnson GR, Ristovski ZD, Hargreaves M, Mengersen K, Corbett S, et al. Size distribution and sites of origin of droplets expelled from the human respiratory tract during expiratory activities. J Aerosol Sci. 2009;40(3):256–69.

44. Wurie FB, Lawn SD, Booth H, Sonnenberg P, Hayward AC. Bio-aerosol production by patients with tuberculosis during normal tidal breathing : implications for transmission risk : a cohort study. Lancet. Elsevier Ltd;

386:S81.

45. Sultan L, Nyka W, Mills C, O'Grady F, Wells W, Riley RL. Tuberculosis disseminators. A study of the variability of aerial infectivity of tuberculous patients. Am Rev Respir Dis. 1960;82:358–69.

46. Riley RL, Mills CC, O'Grady F, Sultan LU, Wittstadt F, Shivpuri DN. Infectiousness of air from a tuberculosis ward. Ultraviolet irradiation of infected air: comparative infectiousness of different patients. Am Rev Respir Dis. 1962;85:511–25.

47. Escombe AR, Oeser C, Gilman RH, Navincopa M, Ticona E, Martínez C, et al. The detection of airborne transmission of tuberculosis from HIV-infected patients, using an in vivo air sampling model. Clin Infect Dis. 2007;44:1349–57.

48. Escombe AR, Moore DAJ, Gilman RH, Pan W, Navincopa M, Ticona E, et al. The infectiousness of tuberculosis patients coinfected with HIV. PLoS Med. 2008 Sep 30;5(9):e188.

49. Dharmadhikari AS, Mphahlele M, Venter K, Stoltz A, Mathebula R, Masotla T, et al. Rapid impact of effective treatment on transmission of multidrug-resistant tuberculosis. Int J Tuberc Lung Dis. 2014;18(9):1019–25.

50. Kamat SR, Dawson JJY, Devadatta S, Fox W, Janardhanam B, Radhakrishna S, et al. A Controlled Study of the Influence of Segregation of Tuberculous Patients for One Year on the Attack Rate of Tuberculosis in a 5-Year Period in Close Family Contacts in South India. Bull World Heal Organ. 1966;34:517–32.

51. Menzies D. Effect of Treatment on Contagiousness of Patients With Active Pulmonary Tuberculosis. Infect Control Hosp Epidemiol. 1997;18(8):582–6.

52. Riley EC, Murphy G, Riley RL. Airborne Spread of Measles in a Suburban

Elementary School. Am J Epidemiol. 1978;107(5):421–32.

53. Corbett EL, Charalambous S, Moloi VM, Fielding K, Grant AD, Dye C, et al. Human immunodeficiency virus and the prevalence of undiagnosed tuberculosis in African gold miners. Am J Respir Crit Care Med. 2004 Sep 15;170(6):673–9.

54. Corbett EL, Bandason T, Yin BC, Munyati S, Godfrey-Faussett P, Hayes R, et al. Epidemiology of tuberculosis in a high HIV prevalence population provided with enhanced diagnosis of symptomatic disease. PLoS Med. 2007;4(1):e22.

55. Adams WC. How much air do we breathe? [Internet]. Sacramento; 1994. Available from: http://www.arb.ca.gov/research/resnotes/notes/94-11.htm

56. Riley R, Mills C, Nyka W, Weinstock N, Storey P, Sultan L, et al. Aerial dissemination of pulmonary tuberculosis. Am J Hyg. 1959;70:185–96.

57. Riley RL, Nardell EA. Clearing the air. The theory and application of ultraviolet air disinfection. Am Rev Respir Dis. 1989;139:1286–94.

58. Dharmadhikari AS, Mphahlele M, Stoltz A, Venter K, Mathebula R, Masotla T, et al. Surgical face masks worn by patients with multidrug-resistant tuberculosis: Impact on infectivity of air on a hospital ward. Am J Respir Crit Care Med. 2012;185(10):1104–9.

59. Beggs CB, Noakes CJ, Sleigh PA, Fletcher LA, Siddiqi K. The transmission of tuberculosis in confined spaces: An analytical review of alternative epidemiological models. Int J Tuberc Lung Dis. 2003;7(11):1015–26.

60. Andrews JR, Morrow C, Walensky RP, Wood R. Integrating Social Contact and Environmental Data in Evaluating Tuberculosis Transmission in a South African Township. J Infect Dis. 2014 Apr 28;210(4):597–603.

61. Nardell EA, Keegan J, Cheney SA, C. ES. Airborne Infection. Theoretical limits of protection achievable by building ventilation. Am Rev Respir Dis.

1991;144:302–6.

62. Godfrey-Faussett P, Sonnenberg P, Shearer SC, Bruce MC, Mee C, Morris L, et al. Tuberculosis control and molecular epidemiology in a South African gold-mining community. Lancet. 2000;356:1066–71.

63. Ypma RJF, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology. 2013 May;24(3):395–400.

64. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis. Elsevier Ltd; 2013 Feb;13(2):137–46.

65. Shaw J, Wynn-Williams N. Infectivity of pulmonary tuberculosis in relation to sputum status. Am Rev Tuberc. 1954;69(5):724–32.

66. Grzybowski S, Barnett G, Styblo K. Contacts of cases of active pulmonary tuberculosis. Bull Int Union Tuberc. 1975;50(1):90–106.

67. van Geuns H, Meijer J, Styblo K. Results of contact examination in Rotterdam, 1967-1969. Bull Int Union Tuberc. 1975;50(1):107–21.

68. Munthali L, Khan PY, Mwaungulu NJ, Chilongo F, Floyd S, Kayange M, et al. The effect of HIV and antiretroviral therapy on characteristics of pulmonary tuberculosis in northern Malawi: a cross-sectional study. BMC Infect Dis. 2014;14:107.

69. van Halsema CL, Fielding KL, Chihota VN, George EC, Lewis JJ, Churchyard GJ, et al. Brief Report: The Effect of Antiretroviral Therapy and CD4 Count on Markers of Infectiousness in HIV-Associated Tuberculosis. JAIDS. 2015;70(1):104–8.

70. Johnstone-robertson SP, Mark D, Morrow C, Middelkoop K, Chiswell M,

Aquino LDH, et al. Social Mixing Patterns Within a South African Township Community : Implications for Respiratory Disease Transmission and Control. 2011;174(11):1246–55.

71. Dodd PJ, Looker C, Plumb ID, Bond V, Schaap A, Shanaube K, et al. Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection. Am J Epidemiol. 2016;183(2):156–66.

72. McCreesh N, Looker C, Dodd PJ, Plumb ID, Shanaube K, Muyoyeta M, et al. Comparison of indoor contact time data in Zambia and Western Cape, South Africa suggests targeting of interventions to reduce Mycobacterium tuberculosis transmission should be informed by local data. BMC Infect Dis. BMC Infectious Diseases; 2016;16(1):71.

73. Barrera E, Livchits V, Nardell E. F-A-S-T : a refocused , intensified , administrative tuberculosis transmission control strategy. Int J Tuberc Lung Dis. 2015;19(January):381–4.

74. Li Y, Leung GM, Tang JW, Yang X, Chao CYH, Lin JZ, et al. Role of ventilation in airborne transmission of infectious agents in the built environment ? a multidisciplinary systematic review. Indoor Air. 2007;17(1):2–18.

75. Chamie G, Wandera B, Luetkemeyer A, Bogere J, Mugerwa RD, Havlir D V, et al. Household ventilation and tuberculosis transmission in Kampala, Uganda. Int J Tuberc Lung Dis. International Union Against Tuberculosis and Lung Disease; 2013 Jun;17(6):764–70.

76. Escombe AR, Moore DAJ, Gilman RH, Navincopa M, Ticona E, Sheen P, et al. Upper-Room Ultraviolet Light and Negative Air Ionization to Prevent Tuberculosis Transmission. PLoS Med. 2009;6(3):e1000043.

77. Verver S, Warren RM, Munch Z, Richardson M, van der Spuy GD, Borgdorff MW, et al. Proportion of tuberculosis transmission that takes place in

households in a high-incidence area. Lancet. 2004 Jan 17;363(9404):212–4.

78. Horna-Campos OJ, Sanchez-Perez HJ, Sanchez I, Bedoya A, Martín M. Public Transportation and Pulmonary Tuberculosis, Lima, Peru. Emerg Infect Dis. 2007;13(10):1491–3.

79. Horna-Campos OJ, Bedoya-Lama A, Romero-Sandoval NC, Martin-Mateo M. Risk of tuberculosis in public transport sector workers , Lima , Peru. Int J Tuberc Lung Dis. 2010;14(September 2009):714–9.

80. Horna-Campos OJ, Consiglio E, Sánchez-Pérez HJ, Navarro A, Caylà J a, Martín-Mateo M. Pulmonary tuberculosis infection among workers in the informal public transport sector in Lima, Peru. Occup Environ Med. 2011 Feb;68(2):163–5.

81. Jackson C, Mostowy JH, Stagg HR, Abubakar I, Andrews N, Yates TA. Working conditions and tuberculosis mortality in England and Wales, 1890–1912: a retrospective analysis of routinely collected data. BMC Infect Dis. BMC Infectious Diseases; 2016;16(1):215.

82. Corbett EL, Churchyard GJ, Clayton TC, Williams BG, Mulder D, Hayes RJ, et al. HIV infection and silicosis: the impact of two potent risk factors on the incidence of mycobacterial disease in South African miners. AIDS. 2000 Dec;14(17):2759–68.

83. Baussano I, Nunn P, Williams B, Pivetta E, Bugiani M, Scano F. Tuberculosis among health care workers. Emerg Infect Dis. 2011 Mar;17(3):488–94.

84. Hanifa Y, Fielding KL, Charalambous S, Variava E, Luke B, Churchyard GJ, et al. Tuberculosis among adults starting antiretroviral therapy in South Africa: The need for routine case finding. Int J Tuberc Lung Dis. 2012;16(9):1252–9.

85. Claassens MM, Jacobs E, Cyster E, Jennings K, James A, Dunbar R, et al. Tuberculosis cases missed in primary health care facilities: should we redefine

case finding? Int J Tuberc Lung Dis. 2013;17(5):608–14.

86. Gandhi NR, Weissman D, Moodley P, Ramathal M, Elson I, Kreiswirth BN, et al. Nosocomial transmission of extensively drug-resistant tuberculosis in a rural hospital in South Africa. J Infect Dis. 2013 Jan 1;207(1):9–17.

87. Bantubani N, Kabera G, Connolly C, Rustomjee R, Reddy T, Cohen T, et al. High rates of potentially infectious tuberculosis and multidrug-resistant tuberculosis (MDR-TB) among hospital inpatients in KwaZulu Natal, South Africa indicate risk of nosocomial transmission. PLoS One. 2014;9(3):1–7.

88. South African Census Bureau. Statistical release ( Revised ) Census 2011 [Internet]. 2012. Available from: www.statssa.gov.za

89. Massyn N, Peer N, Padarath A, Barron P, C D. District Health Barometer 2014/15. Durban; 2015.

90. Dilraj A, Bristow CC, Connolly C, Margot B, Dlamini S, Podewils LJ. Validation of sputum smear results in the Electronic TB Register for the management of tuberculosis , South Africa. Int J Tuberc Lung Dis. 2013;17(November 2012):1317–21.

91. Bristow CC, Dilraj A, Margot B, Jean L. Lack of patient registration in the electronic TB register for sputum smear-positive patients in KwaZulu-Natal, South Africa. Tuberculosis (Edinb). Elsevier Ltd; 2013 Sep;93(5):567–8.

92. Nanoo A, Izu A, Ismail NA, Ihekweazu C, Abubakar I, Mametja D, et al. Nationwide and regional incidence of microbiologically confirmed pulmonary tuberculosis in South Africa, 2004–12: a time series analysis. Lancet Infect Dis. Elsevier Ltd; 2015;3099(15):1–11.

93. Shisana, O, Rehle, T, Simbayi LC, Zuma, K, Jooste, S, Zungu N, Labadarios, D, Onoya D. South African National HIV Prevalence, Incidence and Behaviour Survey, 2012. HSRC Press. 2014.

94. Zaidi J, Grapsa E, Tanser F, Newell M-L, Bärnighausen T. Dramatic increase in HIV prevalence after scale-up of antiretroviral treatment. AIDS. 2013 Sep 10;27(14):2301–5.

95. Wallengren K, Scano F, Nunn P, Margot B, Buthelezi S, Williams B, et al. Resistance to TB drugs in KwaZulu-Natal : causes and prospects for control. 2007. Report No.: arXiv:1107.1800 [q-bio.CB].

96. Arabin G, Gärtig D, Kleeberg HH. First tuberculosis prevalence survey in KwaZulu. S Afr Med J. 1979;56(11):434–8.

97. Richardson ET, Morrow CD, Ho T, Fürst N, Cohelia R, Tram KH, et al. Forced removals embodied as tuberculosis. Soc Sci Med. Elsevier Ltd; 2016;161:13–8.

98. World Health Organisation. WHO TB Burden Estimates [Internet]. 2015 [cited 2015 Sep 4]. Available from: http://www.who.int/tb/country/data/download/en/

99. Barnes PF, Yang Z, Pogoda JM, Preston-Martin S, Jones BE, Otaya M, et al. Foci of tuberculosis transmission in central Los Angeles. Am J Respir Crit Care Med. 1999 Apr;159(4 Pt 1):1081–6.

100. Baussano I, Williams BG, Nunn P, Beggiato M, Fedeli U, Scano F. Tuberculosis Incidence in Prisons: A Systematic Review. PLoS Med. 2010;7(12):e1000381.

101. Van Helden PD. Molecular Epidemiology of TB : Challenging Dogmas and Asking New Questions. IUBMB Life. 2002;53:219–23.

102. Borgdorff MW, van Soolingen D. The re-emergence of tuberculosis: what have we learnt from molecular epidemiology? Clin Microbiol Infect. 2013 Oct 4;19(10):889–901.

103. van Embden J, Cave MD, Crawford JT, Dale JW, Eisenach KD, Gicquel B, et al. Strain Identification of Mycobacterium tuberculosis by DNA Fingerprinting :

Recommendations for a Standardized Methodology. J Clin Microbiol. 1993;31(2):406–9.

104. Glynn JR, Bauer J, De Boer a. S, Borgdorff MW, Fine PEM, Godfrey-Faussett P, et al. Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis. Int J Tuberc Lung Dis. 1999;3(12):1055–60.

105. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. Mol Microbiol. 2000 May;36(3):762–71.

106. Supply P, Lesjean S, Savine E, Soolingen D Van, Locht C, Kremer K, et al. Automated High-Throughput Genotyping for Study of Global Epidemiology of Mycobacterium tuberculosis Based on Mycobacterial Interspersed Repetitive Units. J Clin Microbiol. 2001;39(10):3563–71.

107. Supply P, Allix C, Lesjean S, Cardoso-Oelemann M, Rüsch-Gerdes S, Willery E, et al. Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of Mycobacterium tuberculosis. J Clin Microbiol. 2006 Dec;44(12):4498–510.

108. Roetzer A, Diel R, Kohl T a, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. PLoS Med. 2013 Jan;10(2):e1001387.

109. Bryant JM, Schürch AC, van Deutekom H, Harris SR, de Beer JL, de Jager V, et al. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. BMC Infect Dis. 2013 Feb 27;13(1):110.

110. Guerra-Assunção JA, Crampin AA, Houben RMGJ, Mzembe T, Mallard K, Coll F, et al. Large scale population-based whole genome sequencing of Mycobacterium tuberculosis provides insights into transmission in a high

prevalence area. Elife. 2015;4:e05166.

111. Pérez-Lago L, Herranz M, Lirola MM, Bouza E, García de Viedma D. Characterization of microevolution events in Mycobacterium tuberculosis strains involved in recent transmission clusters. J Clin Microbiol. 2011 Nov;49(11):3771–6.

112. Murray M, Alland D. Methodological Problems in the Molecular Epidemiology of Tuberculosis. 2002;155(6):565–71.

113. Borgdorff MW, Van Den Hof S, Kalisvaart N, Kremer K, Van Soolingen D. Influence of sampling on clustering and associations with risk factors in the molecular epidemiology of tuberculosis. Am J Epidemiol. 2011;174(2):243–51.

114. Barendregt JJ, Doi S a, Lee YY, Norman RE, Vos T. Meta-analysis of prevalence. J Epidemiol Community Health. 2013 Nov 1;67(11):974–8.

115. Freeman M, Tukey J. Transformations Related to the Angular and the Square Root. Ann Math Stat. 1950;21(4):607–11.

116. Crampin AC, Glynn JR, Traore H, Yates MD, Mwaungulu L, Mwenebabu M, et al. Tuberculosis Transmission Attributable to Close Contacts and HIV Status, Malawi. Emerg Infect Dis. 2006;12(5):729–35.

117. Wilkinson D, Pillay M, Crump J, Lombard C, Davies GR, Sturm AW. Molecular epidemiology and transmission dynamics of Mycobacterium tuberculosis in rural Africa. Trop Med Int Heal. 1997;2(8):747–53.

118. Buu TN, van Soolingen D, Huyen MNT, Lan NNT, Quy HT, Tiemersma EW, et al. Tuberculosis acquired outside of households, rural Vietnam. Emerg Infect Dis. 2010 Sep;16(9):1466–8.

119. Porta M, editor. A Dictionary of Epidemiology. 5th ed. New York: Oxford University Press; 2008. 289 p.

120. Greenland S, O'Rourke K. Meta-Analysis. In: Rothman KJ, Greenland S, Lash

TL, editors. Modern Epidemiology. 3rd ed. Philadephia: Lippincott, Williams and Wilkins; 2008. p. 652–82.

121. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002 Jun 15;21(11):1539–58.

122. Houben RMGJ, Glynn JR. A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis : development of a new tool to aid interpretation. 2009;14(8):892–909.

123. Mears J, Abubakar I, Cohen T, McHugh TD, Sonnenberg P. Effect of study design and setting on tuberculosis clustering estimates using Mycobacterial Interspersed Repetitive Units-Variable Number Tandem Repeats (MIRU-VNTR): a systematic review. BMJ Open. 2015;5:e005636.

124. Broekmans JF, Migliori GB, Rieder HL, Lees J, Ruutu P, Loddenkemper R, et al. European framework for tuberculosis control and elimination in countries with a low incidence. Eur Respir J. 2002;19(4):765–75.

125. European Centre for Disease Prevention and Control. Progressing towards TB elimination. Prevention. Stockholm: ECDC; 2010. 37 p.

126. Hanekom M, Van Der Spuy GD, Gey Van Pittius NC, McEvoy CRE, Hoek KGP, Ndabambi SL, et al. Discordance between mycobacterial interspersed repetitive-unit-variable- number tandem-repeat typing and IS6110 restriction fragment length polymorphism genotyping for analysis of Mycobacterium tuberculosis Beijing strains in a setting of high incidence o. J Clin Microbiol. 2008;46(10):3338–45.

127. Roetzer A, Schuback S, Diel R, Gasau F, Ubben T, di Nauta A, et al. Evaluation of Mycobacterium tuberculosis typing methods in a 4-year study in Schleswig-Holstein, Northern Germany. J Clin Microbiol. 2011 Dec;49(12):4173–8.

128. Glynn JR, Vynnycky E, Fine PE. Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. Am J Epidemiol. 1999;149(4):366–71.

129. Borgdorff MW, Sebek M, Geskus RB, Kremer K, Kalisvaart N, van Soolingen D. The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. Int J Epidemiol. 2011;40(4):964–70.

130. Grandjean L, Gilman RH, Martin L, Soto E, Castro B, Lopez S, et al. Transmission of Multidrug-Resistant and Drug-Susceptible Tuberculosis within Households: A Prospective Cohort Study. PLOS Med. 2015;12(6):e1001843.

131. Giesecke J. Modern Infectious Disease Epidemiology. 2nd ed. London: Arnold; 2002. 268 p.

132. Behr M, Warren S, Salamon H, Hopewell PC, Ponce de Leon A, Daley CL, et al. Transmission of Mycobacterium tuberculosis from patients smear-negative for acid-fast bacilli. Lancet. 1999 Feb 6;353(9151):444–9.

133. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact investigation for tuberculosis: a systematic review and meta-analysis. Eur Respir J. 2013 Jan;41(1):140–56.

134. Huh YJ, Aim DI, Kim SJ. Limited variation of DNA fingerprints ( IS6110 and IS1081 ) in Korean strains. Tuber Lung Dis. 1995;76:324–9.

135. Classen CN, Warren R, Richardson M, Hauman JH, Gie RP, Ellis JH, et al. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. Thorax. 1999 Feb;54(2):136–40.

136. Teixeira L, Perkins MD, Johnson JL, Keller R, Palaci M, Dettoni VV, et al. Infection and disease among household contacts of patients with multidrug-resistant tuberculosis. Int J Tuberc Lung Dis. 2001;5(January):321–8.

137. Bennett DE, Onorato IM, Ellis BA, Crawford JT, Schable B, Byers R, et al.

DNA Fingerprinting of Mycobacterium tuberculosis Isolates from Epidemiologically Linked Case Pairs. Emerg Infect Dis. 2002;8(11):1224–9.

138. Schaaf HS, Michaelis IA, Richardson M, Booysen CN, Gie RP, Warren R, et al. Adult-to-child transmission of tuberculosis : household or community contact ? Int J Tuberc Lung Dis. 2003;7(August 2002):426–31.

139. Inigo J, Arce A, Martin-Moreno J, Herruzo R, Palenque E, Chaves F. Recent transmission of tuberculosis in Madrid: application of capture-recapture analysis to conventional and molecular epidemiology. Int J Epidemiol. 2003 Oct 1;32(5):763–9.

140. Martín A, Iñigo J, Chaves F, Herranz M, Ruiz-Serrano MJ, Palenque E, et al. Re-analysis of epidemiologically linked tuberculosis cases not supported by IS6110-RFLP-based genotyping. Clin Microbiol Infect. 2009 Aug;15(8):763–9.

141. Borrell S, Español M, Orcau A, Tudó G, March F, Caylà J a, et al. Factors associated with differences between conventional contact tracing and molecular epidemiology in study of tuberculosis transmission and analysis in the city of Barcelona, Spain. J Clin Microbiol. 2009 Jan;47(1):198–204.

142. Nabyonga L, Kateete DP, Katabazi F a, Odong PR, Whalen CC, Dickman KR, et al. Determination of circulating Mycobacterium tuberculosis strains and transmission patterns among pulmonary TB patients in Kawempe municipality, Uganda, using MIRU-VNTR. BMC Res Notes. BioMed Central Ltd; 2011 Jan;4(1):280.

143. Augustynowicz-Kopeć E, Jagielski T, Kozińska M, Kremer K, van Soolingen D, Bielecki J, et al. Transmission of tuberculosis within family-households. J Infect. 2012 Jun;64(6):596–608.

144. Leung ECC, Leung CC, Kam KM, Yew WW, Chang KC, Leung WM, et al. Transmission of multidrug-resistant and extensively drug-resistant tuberculosis

in a metropolitan city. Eur Respir J. 2013 Apr;41(4):901–8.

145. Sia IG, Buckwalter SP, Doerr KA, Lugos S, Kramer R, Orillaza-chi R, et al. Genotypic characteristics of Mycobacterium tuberculosis isolated from household contacts of tuberculosis patients in the Philippines. BMC Infect Dis. 2013;13:571.

146. Behr MA, Hopewell PC, Paz EA, Kawamura LM, Schecter GF, Small PM. Predictive Value of Contact Investigation for Identifying Recent Transmission of Mycobacterium tuberculosis. Am J Respir Crit Care Med. 1998;158:465–9.

147. Whalen CC, Zalwango S, Chiunda A, Malone L, Eisenach K, Joloba M, et al. Secondary attack rate of tuberculosis in urban households in Kampala, Uganda. PLoS One. 2011 Jan;6(2):e16137.

148. Middelkoop K, Mathema B, Myer L, Shashkina E, Whitelaw A, Kaplan G, et al. Transmission of Tuberculosis in a South African Community With a High Prevalence of HIV Infection. J Infect Dis. 2015 Jul 22;211(1):53–61.

149. Glynn JR, Guerra-Assunção JA, Houben RMGJ, Sichali L, Mzembe T, Mwaungulu LK, et al. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. PLoS One. 2015;10(7):e0132840.

150. Van Wyk SS, Mandalakas a. M, Enarson D a., Gie RP, Beyers N, Hesseling a. C. Tuberculosis contact investigation in a high-burden setting: House or household? Int J Tuberc Lung Dis. 2012;16(2):157–62.

151. Crampin AC, Glynn JR, Fine PEM. What has " Karonga " taught us? Tuberculosis studied over three decades. Int J Tuberc Lung Dis. 2012;13(2):153–64.

152. Story A, Murad S, Roberts W, Verheyen M, Hayward AC. Tuberculosis in London: the importance of homelessness, problem drug use and prison.

Thorax. 2007;62(8):667–71.

153. Narain R, Nair SS, Rao GR, Chandrasekhar P. Distribution of tuberculous infection and disease among households in a rural community. Bull World Health Organ. 1966;34(4):639–54.

154. Madico G, Gilman RH, Checkley W, Cabrera L, Kohlstadt I, Kacena K, et al. Community infection ratio as an indicator for tuberculosis control. Lancet. 1995;345(8947):416–9.

155. Middelkoop K, Bekker L-G, Morrow C, Lee N, Wood R. Decreasing household contribution to TB transmission with age: a retrospective geographic analysis of young people in a South African township. BMC Infect Dis. 2014 Jan;14(1):221.

156. Wood R, Racow K, Bekker L-G, Morrow C, Middelkoop K, Mark D, et al. Indoor social networks in a South African township: potential contribution of location to tuberculosis transmission. PLoS One. 2012 Jan;7(6):e39246.

157. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med. 2008 Mar 25;5(3):e74.

158. Odone A, Crampin AC, Mwinuka V, Malema S, Mwaungulu JN, Munthali L, et al. Association between socioeconomic position and tuberculosis in a large population-based study in rural Malawi. PLoS One. 2013 Jan;8(10):e77740.

159. Wood R, Morrow C, Ginsberg S, Piccoli E, Kalil D, Sassi A, et al. Quantification of shared air: a social and environmental determinant of airborne disease transmission. PLoS One. 2014 Jan;9(9):e106622.

160. Rudnick SN, Milton DK. Risk of indoor airborne infection transmission estimated from carbon dioxide concentration. Indoor Air. 2003 Sep;13(3):237–45.

161. Lillebaek T, Dirksen A, Baess I, Strunge B, Thomsen VØ, Andersen AB. Molecular evidence of endogenous reactivation of Mycobacterium tuberculosis after 33 years of latent infection. J Infect Dis. 2002;185(3):401–4.

162. Cohen T, van Helden PD, Wilson D, Colijn C, McLaughlin MM, Abubakar I, et al. Mixed-strain Mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. Clin Microbiol Rev. 2012;25(4):708–19.

163. Burman WJ, Reves RR. Review of false-positive cultures for Mycobacterium tuberculosis and recommendations for avoiding unnecessary treatment. Clin Infect Dis. 2000;31(6):1390–5.

164. Doi SAR, Barendregt JJ, Khan S, Thalib L, Williams GM. Advances in the meta-analysis of heterogeneous clinical trials I: The inverse variance heterogeneity model. Contemp Clin Trials. Elsevier Inc.; 2015;45:130–8.

165. Joshi R, Reingold AL, Menzies D, Pai M. Tuberculosis among health-care workers in low- and middle-income countries: a systematic review. PLoS Med. 2006 Dec;3(12):e494.

166. Basu S, Stuckler D, Gonsalves G, Lurie M. Globalization and Health The production of consumption : addressing the impact of mineral mining on tuberculosis in southern Africa. Global Health. 2009;5:11.

167. Houlihan CF, Bland RM, Mutevedzi PC, Lessells RJ, Ndirangu J, Thulare H, et al. Cohort Profile : Hlabisa HIV Treatment and Care Programme. Int J Epidemiol. 2011;40(February 2010):318–26.

168. Massyn N, Day C, Peer N, Padarath A, Barron P, English R. District Health Barometer 2013/14. Durban; 2014.

169. Tanser F, Gijsbertsen B, Herbst K. Modelling and understanding primary health care accessibility and utilization in rural South Africa : An exploration using a geographical information system. Soc Sci Med. 2006;63:691–705.

170. Cooke GS, Tanser FC, Bärnighausen TW, Newell M-L. Population uptake of antiretroviral treatment through primary care in rural South Africa. BMC Public Health. 2010;10(1):585.

171. Bor J, Herbst AJ, Newell M-L, Bärnighausen T. Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. Science. 2013 Feb 22;339(6122):961–5.

172. Wallrauch C, Heller T, Lessells R, Kekana M, Barnighausen T, Newell M-L. High uptake of HIV testing for tuberculosis patients in an integrated primary health care HIV/TB programme in rural KwaZulu-Natal. S Afr Med J. 2010 Mar;100(3):146–7.

173. Mossong J, Byass P, Herbst K. Who died of what in rural KwaZulu-Natal, South Africa: a cause of death analysis using InterVA-4. Glob Health Action. 2014;7:25496.

174. Rieder H. Annual risk of infection with Mycobacterium tuberculosis. Eur Repir J. 2005;25(1):181–5.

175. Bowerman RJ. Tuberculin skin testing in BCG-vaccinated populations of adults and children at high risk for tuberculosis in Taiwan. Int J Tuberc Lung Dis. 2004;8(10):1228–33.

176. Fourie P. The prevalence and annual rate of tuberculosis infection in South Africa. Tubercle. 1983;64(3):181–92.

177. Styblo K, Muwinge H, Chum H, Sutherland M, Bleiker M, Broekmans J, et al. The second round of the national tuberculin survey in Tanzania, 1988–1992. 1995.

178. Arnadottir T, Soukaseum H, Vangvichit P, Bounmala S, Vos E. Prevalence and annual risk of tuberculosis infection in laos. Int J Tuberc Lung Dis. 2001;5(5):391–9.

179. Trebucq A, Guerin N, Ali Ismael H, Bernatas JJ, Sevre JP, Rieder HL. Prevalence and trends of infection with Mycobacterium tuberculosis in Djibouti, testing an alternative method. Int J Tuberc Lung Dis. 2005;9(10):1097–104.

180. Tanzania Tuberculin Survey Collaboration. Tuberculosis control in the era of the HIV epidemic: Risk of tuberculosis infection in Tanzania, 1983-1998. Int J Tuberc Lung Dis. 2001;5(2):103–12.

181. Benaglia T, Chauveau D, Hunter DR, Young DS. mixtools: An R Package for Analyzing Finite. J Stat Softw. 2009;32(6):1–29.

182. Young D, Benaglia T, Chauveau D, Hunter D, Elmore R, Hettmansperger T, et al. mixtools: Tools for Analyzing Finite Mixture Models. 2016.

183. Neuenschwander B. Bayesian Mixture Analysis for Tuberculin Induration Data [Internet]. 2007. p. 82. Available from: http://www.tbrieder.org/research/mixture/mixture.html

184. Davies GR, Fine PE, Vynnycky E. Mixture analysis of tuberculin survey data from northern Malawi and critique of the method. Int J Tuberc Lung Dis. 2006;10(9):1023–9.

185. Shanaube K, Sismanidis C, Ayles H, Beyers N, Schaap A, Lawrence A, et al. Annual Risk of Tuberculous Infection Using Different Methods in Communities with a High Prevalence of TB and HIV in Zambia and South Africa. PLoS One. 2009;4(11).

186. Addo KK, van den Hof S, Mensah GI, Hesse A, Bonsu C, Koram KA, et al. A tuberculin skin test survey among Ghanaian school children. BMC Public Health. 2010 Jan;10:35.

187. Pai M, Dendukuri N, Wang L, Joshi R, Kalantri S, Rieder HL. Improving the estimation of tuberculosis infection prevalence using T-cell based assay and mixture models. Int J Tuberc Lung Dis. 2008;12(January):895–902.

188. Aggerbeck H, Giemza R, Joshi P, Tingskov PN, Hoff ST, Boyle J, et al. Randomised clinical trial investigating the specificity of a novel skin test (C-Tb) for diagnosis of M. tuberculosis infection. PLoS One. 2013 Jan;8(5):e64215.

189. Hoff ST, Peter JG, Theron G, Pascoe M, Tingskov PN, Aggerbeck H, et al. Sensitivity of C-Tb: A novel RD-1-specific skin test for the diagnosis of tuberculosis infection. Eur Respir J. 2016;47(3):919–28.

190. Nyboe J. Interpretation of tuberculosis infection age curves. Bull World Heal Organ. 1957;17:319–39.

191. Pretorius C, Bacaer N, Williams B, Wood R, Ouifki R. On the Relationship between Age, Annual Rate of Infection, and Prevalence of Mycobacterium tuberculosis in a South African Township. Clin Infect Dis. 2009;48:994–6.

192. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. Stat Med. 1995;14(December 1993):799–810.

193. Crofton J. The Battle with the Bug. Saving lives and preventing misery - The memoirs of Professor Sir John Wenman Crofton [Internet]. 1993. p. 339–460. Available from: https://www.rcpe.ac.uk/library-archives/sir-john-crofton-autobiography

194. Kwamanga D, Chakaya J, Sitienei J, Kalisvaart N, Herminez RL, Werf MJ Van Der. Tuberculosis transmission in Kenya : results of the third National Tuberculin Survey. 2010;14(August 2009):695–700.

195. Tanser F, Bärnighausen T, Cooke GS, Newell M-L. Localized spatial clustering of HIV infections in a widely disseminated rural South African epidemic. Int J Epidemiol. 2009 Aug;38(4):1008–16.

196. Hunter JM, Arbona S. Field testing along a disease gradient: Some geographical dimensions of tuberculosis in Puerto Rico. Soc Sci Med. 1985;21(9):1023–42.

197. Khan PY, Glynn JR, Fielding KL, Mzembe T, Mulawa D, Chiumya R, et al. Risk factors for Mycobacterium tuberculosis infection in 2 – 4 year olds in a rural HIV-prevalent setting. Int J Tuberc Lung Dis. 2016;20(3):342–9.

198. Middelkoop K, Bekker L-G, Morrow C, Zwane E, Wood R. Childhood tuberculosis infection and disease: a spatial and temporal transmission analysis in a South African township. South African Med J. 2009;99(10):738–43.

199. Zelner JL, Murray MB, Becerra MC, Galea J, Lecca L, Calderon R, et al. Identifying Hotspots of Multidrug-Resistant Tuberculosis Transmission Using Spatial and Molecular Genetic Data. J Infect Dis. 2016;213:287–94.

200. Jenkins HE, Gegia M, Furin J, Kalandadze I, Nanava U, Chakhaia T, et al. Geographical heterogeneity of multidrug-resistant tuberculosis in Georgia, January 2009 to June 2011. Eurosurveillance. 2014;19(11):1–10.

201. Jenkins HE, Plesca V, Ciobanu A, Crudu V, Galusca I, Soltan V, et al. Assessing spatial heterogeneity of multidrug-resistant tuberculosis in a high-burden country. Eur Respir J. 2013;42(5):1291–301.

202. van der Werf M. Generic protocol for tuberculin school survey [Internet]. The Hague; 2007. Available from: http://www.kncvtbc.nl

203. Vyas S, Kumaranayake L. Constructing socio-economic status indices: how to use principal components analysis. Health Policy Plan. 2006 Nov;21(6):459–68.

204. Filmer D, Pritchett L. Estimating wealth effects without expenditure data — or tears: an application to educational enrollments in states of India. Demography. 2001;38(1):115–32.

205. Falkingham J, Namazie C. Measuring health and poverty : a review of approaches to identifying the poor. London; 2002.

206. Howe LD, Galobardes B, Matijasevich A, Gordon D, Johnston D, Onwujekwe O, et al. Measuring socio-economic position for epidemiological studies in low- and middle-income countries: a methods of measurement in epidemiology paper. Int J Epidemiol. 2012 Jun;41(3):871–86.

207. Hosegood V, Floyd S, Marston M, Hill C, McGrath N, Isingo R, et al. The effects of high HIV prevalence on orphanhood and living arrangements of children in Malawi, Tanzania, and South Africa. Popul Stud (NY). 2007;61(3):327–36.

208. Martin Kulldorff and Information Management Services Inc. SaTScan [Internet]. 2014. Available from: www.satscan.org

209. Mahomed H, Hawkridge T, Verver S, Geiter L, Hatherill M, Abrahams D, et al. Predictive factors for latent tuberculosis infection among adolescents in a high-burden area in South Africa. Int J Tuberc Lung Dis. 2011 Mar;15(3):331–6.

210. Middelkoop K, Bekker L, Myer L, Dawson R, Wood R. Rates of Tuberculosis Transmission to Children and Adolescents in a Community with a High Prevalence of HIV Infection among Adults. 2008;7705:349–55.

211. Middelkoop K, Bekker L-G, Liang H, Aquino LDH, Sebastian E, Myer L, et al. Force of tuberculosis infection among adolescents in a high HIV and TB prevalence community: a cross-sectional observation study. BMC Infect Dis. 2011;11(1):156.

212. Eilers PHC, Borgdorff MW. Modeling and correction of digit preference in tuberculin surveys. Int J Tuberc Lung Dis. 2004;8(2):232–9.

213. Cauthen GM, Pio A, ten Dam HG. Annual risk of tuberculous infection. Geneva; 1988.

214. Vynnycky E, Fine PEM. The annual risk of infection with Mycobacterium tuberculosis in England and Wales since 1901. Int J Tuberc Lung Dis.

1997;1(5):389–96.

215. Sutherland I, Bleiker MA, Meijer J, Styblo K. The risk of tuberculous infection in The Netherlands from 1967 to 1979. Tubercle. 1983;64(4):241–53.

216. Feske ML, Teeter LD, Musser JM, Graviss E a. Including the third dimension: a spatial analysis of TB cases in Houston Harris County. Tuberculosis (Edinb). Elsevier Ltd; 2011 Dec;91 Suppl 1:S24–33.

217. Furukawa NW, Mendoza-Ticona A, Alarcon-Villaverde JO, Montejo H, Micek MA, Zunt JR. The association between public transport and active tuberculosis in Lima, Peru. Eur Respir J. 2014;43:1192–5.

218. Murray EJ, Marais BJ, Mans G, Beyers N, Ayles H, Godfrey-Faussett P, et al. A multidisciplinary method to map potential tuberculosis transmission “hot spots” in high-burden communities. International Union Against Tuberculosis and Lung Disease;

219. Wood R, Liang H, Wu H, Middelkoop K, Oni T, Rangaka MX, et al. Changing prevalence of tuberculosis infection with increasing age in high-burden townships in South Africa. Int J Tuberc Lung Dis. 2010 Apr;14(4):406–12.

220. Khan P, Mzembe T, Kranzer K, Mulawa D, Koole O, Fielding K, et al. PC-1173-06 Mycobacterium tuberculosis transmission is associated with attendance at community gathering places in rural Malawi. Int J Tuberc Lung Dis. 2015;19(12 (supplement 2)):S483–4.

221. WHO. Stop TB Strategy. Building on and enhancing DOTS to meet the TB-related Millenium Development Goals. Geneva; 2006.

222. Corbett EL, Charalambous S, Fielding K, Clayton T, Hayes RJ, De Cock KM, et al. Stable incidence rates of tuberculosis (TB) among human immunodeficiency virus (HIV)-negative South African gold miners during a decade of epidemic HIV-associated TB. J Infect Dis. 2003;188(8):1156–63.

223. Suthar AB, Lawn SD, del Amo J, Getahun H, Dye C, Sculier D, et al. Antiretroviral therapy for prevention of tuberculosis in adults with HIV: a systematic review and meta-analysis. PLoS Med. 2012 Jan;9(7):e1001270.

224. Dodd PJ, Knight GM, Lawn SD, Corbett EL, White RG. Predicting the Long-Term Impact of Antiretroviral Therapy Scale-Up on Population Incidence of Tuberculosis. PLoS One. 2013;8(9):e75466.

225. Hossain S, Quaiyum MA, Zaman K, Banu S, Husain MA, Islam MA, et al. Socio economic position in TB prevalence and access to services: results from a population prevalence survey and a facility-based survey in Bangladesh. PLoS One. 2012 Jan;7(9):e44980.

226. Zaman K, Hossain S, Banu S, Quaiyum M a, Barua PC, Salim M a H, et al. Prevalence of smear-positive tuberculosis in persons aged ≥ 15 years in Bangladesh: results from a national survey, 2007-2009. Epidemiol Infect. 2012 Jun;140(6):1018–27.

227. van Leth F, Guilatco RS, Hossain S, Van't Hoog AH, Hoa NB, van der Werf MJ, et al. Measuring socio-economic data in tuberculosis prevalence surveys. Int J Tuberc Lung Dis. 2011 Jun;15 Suppl 2:S58–63.

228. Corbett EL, Bandason T, Cheung Y-B, Makamure B, Dauya E, Munyati SS, et al. Prevalent infectious tuberculosis in Harare, Zimbabwe: burden, risk factors and implications for control. Int J Tuberc Lung Dis. 2009 Oct;13(10):1231–7.

229. Montagu D, Sudhinaraset M, Lwin T, Onozaki I, Win Z, Aung T. Equity and the Sun Quality Health Private Provider Social Franchise: comparative analysis of patient survey data and a nationally representative TB prevalence survey. Int J Equity Health. 2013 Jan;12:5.

230. Boccia D, Hargreaves J, Stavola BL De, Fielding K, Schaap A, Godfrey-faussett P, et al. The Association between Household Socioeconomic Position

and Prevalent Tuberculosis in Zambia : A Case-Control Study. PLoS One. 2011;6(6):e20824.

231. Boccia D, Hargreaves J, Ayles H, Fielding K, Simwinga M, Godfrey-Faussett P. Tuberculosis infection in Zambia: the association with relative wealth. Am J Trop Med Hyg. 2009 Jun;80(6):1004–11.

232. Lygizos M, Shenoi S V, Brooks RP, Bhushan A, Brust JCM, Zelterman D, et al. Natural ventilation reduces high TB transmission risk in traditional homes in rural KwaZulu-Natal, South Africa. BMC Infect Dis. BMC Infectious Diseases; 2013;13(1):300.

233. Chimbindi N, Bor J, Newell M-L, Tanser F, Baltussen R, Hontelez J, et al. Time and money: The true costs of health care utilization for patients receiving “free” HIV/tuberculosis care and treatment in rural KwaZulu-natal. J Acquir Immune Defic Syndr. 2015;70(2):e52–60.

234. Muniyandi M, Ramachandran R, Gopi PG, Chandrasekaran V, Subramani R, Sadacharam K, et al. The prevalence of tuberculosis in different economic strata: a community survey from South India. Int J Tuberc Lung Dis. 2007 Sep;11(9):1042–5.

235. Hargreaves JR, Bonell CP, Boler T, Boccia D, Birdthistle I, Fletcher A, et al. Systematic review exploring time trends in the association between educational attainment and risk of HIV infection in sub-Saharan Africa. AIDS. 2008 Jan 30;22(3):403–14.

236. Rossi RE, Mulla DJ, Journel AG, Franz EH. Geostatistical Tools for Modeling and Interpreting Ecological Spatial Dependence. Ecol Monogr. 1992;62(2):277–314.

237. Openshaw S. The Modifiable Areal Unit Problem. Norwich, England: Geo Books; 1984.

238. Larmarange J, Mossong J, Barnighausen T, Newell ML. Participation dynamics in population-based longitudinal HIV surveillance in rural South Africa. PLoS One. 2015;10(4):1–16.

239. Martuzzi M, Elliott P. Estimating the incidence rate ratio in cross-sectional studies using a simple alternative to logistic regression. Ann Epidemiol. 1998;8(1):52–5.

240. Tanser F, Bärnighausen T, Grapsa E, Zaidi J, Newell M-L. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. Science. 2013 Feb 22;339(6122):966–71.

241. Farley JE, Tudor C, Mphahlele M, Franz K, Perrin N a., Dorman S, et al. A national infection control evaluation of drug-resistant tuberculosis hospitals in South Africa. Int J Tuberc Lung Dis. 2012;16(1):82–9.

242. McCarthy KM, Scott LE, Gous N, Tellie M, Venter WDF, Stevens WS, et al. High incidence of latent tuberculous infection among South African health workers : an urgent call for action. Int J Tuberc Lung Dis. 2015;19(October 2014):647–53.

243. Andrews JR, Morrow C, Wood R. Modeling the role of public transportation in sustaining tuberculosis transmission in South Africa. Am J Epidemiol. 2013;177(6):556–61.

244. Ximenes RA de A, Albuquerque M de FPM de, Souza W V., Montarroyos UR, Diniz GTN, Luna CF, et al. Is it better to be rich in a poor area or poor in a rich area? A multilevel analysis of a case-control study of social determinants of tuberculosis. Int J Epidemiol. 2009 Oct;38(5):1285–96.

245. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstem AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373–9.

246. Taylor JG, Yates TA, Mthethwa M, Tanser F, Abubakar I, Altamirano H. Measuring ventilation and modelling M . tuberculosis transmission in indoor congregate settings , rural KwaZulu-Natal. 2016;20(9):1155–61.

247. Reniers G, Eaton J. Refusal bias in HIV prevalence estimates from nationally representative seroprevalence surveys. AIDS. 2009 Mar 13;23(5):621–9.

248. Floyd S, Molesworth A, Dube A, Crampin AC, Houben R, Chihana M, et al. Underestimation of HIV prevalence in surveys when some people already know their status, and ways to reduce the bias. AIDS. 2013 Jan 14;27(2):233–42.

249. De Knegt HJ, Van Langevelde F, Coughenour MB, Skidmore AK, De Boer WF, Heitk??nig IMA, et al. Spatial autocorrelation and the scaling of species-environment relationships. Ecology. 2010;91(8):2455–65.

250. De Knegt HJ, Van Langevelde F, Skidmore AK, Delsink A, Slotow R, Henley S, et al. The spatial scaling of habitat selection by African elephants. J Anim Ecol. 2011;80(1):270–81.

251. Egwaga SM, Cobelens FG, Muwinge H, Verhage C, Kalisvaart N, Borgdorff MW. The impact of the HIV epidemic on tuberculosis transmission in Tanzania. AIDS. 2006;20(6):915–21.

252. Richardson ET, Morrow CD, Kalil DB, Bekker L-G, Wood R. Shared air: a renewed focus on ventilation for the prevention of tuberculosis transmission. PLoS One. 2014 Jan;9(5):e96334.

253. Read JM, Edmunds WJ, Riley S, Lessler J, Cummings DAT. Close encounters of the infectious kind: methods to measure social mixing behaviour. Epidemiol Infect. 2012;140(12):2117–30.

254. Barrat A, Cattuto C, Tozzi AE, Vanhems P, Voirin N. Measuring contact patterns with wearable sensors: Methods, data characteristics and

applications to data-driven simulations of infectious diseases. Clin Microbiol Infect [Internet]. European Society of Clinical Infectious Diseases; 2014;20(1):10–6. Available from: http://dx.doi.org/10.1111/1469-0691.12472

255. Kiti MC, Tizzoni M, Kinyanjui TM, Koech DC, Munywoki PK, Meriac M, et al. Quantifying social contacts in a household setting of rural Kenya using wearable proximity sensors. EPJ Data Sci. Kiti et al.; 2016;5(1):21.

256. Ayles H, Muyoyeta M, Du Toit E, Schaap A, Floyd S, Simwinga M, et al. Effect of household and community interventions on the burden of tuberculosis in southern Africa: the ZAMSTAR community-randomised trial. Lancet. Elsevier Ltd; 2013 Oct 5;382(9899):1183–94.

257. Kranzer K, Tomlin K, Golub JE, Shapiro A, Schaap A, Corbett EL, et al. STATE OF THE ART The benefits to communities and individuals of screening for active tuberculosis disease : a systematic review. Int J Tuberc Lung Dis. 2013;17(4):432–46.

258. Ayles HM, Sismanidis C, Beyers N, Hayes RJ, Godfrey-Faussett P. ZAMSTAR, The Zambia South Africa TB and HIV Reduction Study: design of a 2 x 2 factorial community randomized trial. Trials. 2008 Jan;9:63.

259. Kasaie P, Andrews JR, Kelton WD, Dowdy DW. Timing of tuberculosis transmission and the impact of household contact tracing: An agent-based simulation model. Am J Respir Crit Care Med. 2014;189(7):845–52.

260. Rose G. Sick Individuals and Sick Populations. Int J Epidemiol. 1985;14(1):32–8.

261. Nardell EA, Bucher SJ, Brickner PW, Wang C, Vincent RL, Becan-McBride K, et al. Safety of Upper-Room Ultraviolet Germicidal Air Disinfection for Room Occupants: Results from the Tuberculosis Ultraviolet Shelter Study. Public Health Rep. 2008;123(1):52–60.

262. Escombe AR, Oeser CC, Gilman RH, Navincopa M, Ticona E, Pan W, et al. Natural ventilation for the prevention of airborne contagion. PLoS Med. 2007 Feb;4(2):e68.

263. Cox H, Escombe R, Mcdermid C, Mtshemla Y, Spelman T, Azevedo V, et al. Wind-Driven Roof Turbines : A Novel Way to Improve Ventilation for TB Infection Control in Health Facilities. PLoS One. 2012;7(1):e29589.

264. Dowdy D, Chaisson RE. The persistence of tuberculosis in the age of DOTS: reassessing the effect of case detection. Bull World Health Organ. 2009 Apr 1;87(4):296–304.

265. Crawford JT, Braden CR, Schable BA, Onorato IM. National tuberculosis genotyping and surveillance network: Design and methods. Emerg Infect Dis. 2002;8(11):1192–6.

266. Van Rie A, Warren R, Richardson M, Victor TC, Gie RP, Enarson DA, et al. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N Engl J Med. 1999;341:1174–9.

# Appendix 1 – search terms used in the systematic review

## Terms used in MEDLINE and EMBASE

1 exp Tuberculosis/ or exp Mycobacterium tuberculosis/

2 (tuberculos* or TB or Mtb or "M.tb").mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

3 ((mycobacterium or mycobacteria or m or "m.") adj3 (bovis or africanum or microti or canetti)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

4 (tubercle adj3 (bacillus or bacilli)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

5 1 or 2 or 3 or 4

6 (home* or house* or hut* or dwelling* or residence* or flat* or apartment* or domicile* or abode*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

7 ((in or within or between) adj (family or families)).mp.

8 exp Marriage/ or marriage*.mp. or married.mp. or marry*.mp. or husband*.mp. or wife.mp. or wives.mp. or spous*.mp. or cohabit*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

9 ((live or living or stay or staying or reside or residing or sleep or sleeping) adj3 (together or "with each other" or "with eachother")).mp. [mp=title, abstract, original

title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

10 ((close or household or family or prolonged or intimate) adj3 (contact or contacts)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

11 cluster*.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

12 6 or 7 or 8 or 9 or 10 or 11

13 exp Molecular Epidemiology/

14 (molecular adj3 (epidemiology or epidemiological)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

15 exp Disease Transmission, Infectious/ or transmission*.mp. or transmitted.mp. or transmitting.mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

16 exp Bacterial Typing Techniques/

17 ((molecular or strain) adj3 (type or types or typing)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

18 exp Polymorphism, Restriction Fragment Length/

19 (RFLP or "R.F.L.P." or "restriction fragment length polymorphism*").mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

20 IS6110.mp.

21 exp DNA Fingerprinting/ or (fingerprint* or "finger-print*" or "finger print*").mp.

22 exp Interspersed Repetitive Sequences/

23 exp Minisatellite Repeats/

24 (interspersed adj3 repetitive).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

25 ((minisatellite or "mini-satellite" or "mini satellite" or tandem) adj3 repeat*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

26 (MIRU or "M.I.R.U.").mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

27 (VNTR or "V.N.T.R.").mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

28 exp Genotype/

29 exp Sequence Analysis/

30 exp Genome, Bacterial/

31 genotyp*.mp.

32 ((genetic or genome or "whole-genome") adj3 (sequence* or sequencing)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

33 ((genetic or sequence*) adj (analysis or analyses or comparison*)).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]

34 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26 or 27 or 28 or 29 or 30 or 31 or 32 or 33

35 5 and 12 and 34

36 Animals/

37 Humans/

38 36 not 37

39 35 not 38

## Terms used in POPLINE

( ( tuberculos* OR tb OR mtb OR "M.tb" OR bovis OR africanum OR microti OR canetti OR tubercle ) AND ( home* OR house* OR hut* OR dwelling* OR residence* OR flat* OR apartment* OR domicile* OR abode* OR family OR families OR marriage* OR married OR marry* OR husband* OR wife OR wives OR spous* OR cohabit* OR "close contact" OR "household contact" OR "family contact" OR "prolonged contact" OR "intimate contact" OR cluster* ) AND ( "molecular epidemiology" OR "molecular epidemiological" OR transmi* OR "molecular typing" OR "strain typing" OR "RFLP" OR "restriction fragment length polymorphism" OR IS6110 OR fingerprint* OR "finger printing" OR "finger-printing" OR "VNTR" OR "MIRU" OR "MIRU-VNTR" OR genotyp* OR genetic* OR genom* OR "whole-genome" OR sequenc* ) )

## Terms used in Global Health

1 exp Tuberculosis/ or exp Mycobacterium tuberculosis/

2 (tuberculos* or TB or Mtb or "M.tb").mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

3 ((mycobacterium or mycobacteria or m or "m.") adj3 (bovis or africanum or microti or canetti)).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

4 (tubercle adj3 (bacillus or bacilli)).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

5 1 or 2 or 3 or 4

6 (home* or house* or hut* or dwelling* or residence* or flat* or apartment* or domicile* or abode*).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

7 ((in or within or between) adj (family or families)).mp.

8 exp Marriage/ or marriage*.mp. or married.mp. or marry*.mp. or husband*.mp. or wife.mp. or wives.mp. or spous*.mp. or cohabit*.mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

9 ((live or living or stay or staying or reside or residing or sleep or sleeping) adj3 (together or "with each other" or "with eachother")).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

10 ((close or household or family or prolonged or intimate) adj3 (contact or contacts)).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

11 cluster*.mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

12 6 or 7 or 8 or 9 or 10 or 11

13 exp Molecular Epidemiology/

14 (molecular adj3 (epidemiology or epidemiological)).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

15 exp Disease Transmission, Infectious/ or transmission*.mp. or transmitted.mp. or transmitting.mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

16 ((molecular or strain) adj3 (type or types or typing)).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

17 (RFLP or "R.F.L.P." or "restriction fragment length polymorphism*").mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

18 IS6110.mp.

19 exp DNA Fingerprinting/ or (fingerprint* or "finger-print*" or "finger print*").mp.

20 (interspersed adj3 repetitive).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

21 ((minisatellite or "mini-satellite" or "mini satellite" or tandem) adj3 repeat*).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

22 (MIRU or "M.I.R.U.").mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

23 (VNTR or "V.N.T.R.").mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

24 genotyp*.mp.

25 ((genetic or genome or "whole-genome") adj3 (sequence* or sequencing)).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

26 ((genetic or sequence*) adj (analysis or analyses or comparison*)).mp. [mp=abstract, title, original title, broad terms, heading words, identifiers, cabicodes]

27 13 or 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26

28 5 and 12 and 27

29 Animals/

30 Humans/

31 29 not 30

32 28 not 31

## Terms used in Web of Science Conference Proceedings Citations Index

tuberculos* OR tb OR mtb OR "M.tb" OR bovis OR africanum OR microti OR canetti OR tubercle

AND

home* OR house* OR hut* OR dwelling* OR residence* OR flat* OR apartment* OR domicile* OR abode* OR family OR families OR marriage* OR married OR marry* OR husband* OR wife OR wives OR spous* OR cohabit* OR "close contact*" OR "household contact*" OR "family contact*" OR "prolonged contact*" OR "intimate contact*" OR cluster*

AND

“molecular epidemiology” OR “molecular epidemiological” OR transmission* OR transmitted OR transmitting OR “molecular typ*” OR “strain typ*” OR RFLP OR “R.F.L.P.” OR “restriction fragment length polymorphism*” OR IS6110 OR fingerprint* OR “finger print*” OR “finger-print*” OR “interspersed repetitive sequence*” OR “minisatellite repeat*” OR “mini satellite repeat*” OR “mini-satellite repeat*” OR “tandem repeat*” OR MIRU OR “M.I.R.U.” OR VNTR OR “V.N.T.R.” OR genotyp* OR genetic* OR “genom*” OR “whole-genome” OR sequenc*

# Appendix 2 – Summaries of the included studies

## Augustynowicz-Kopec et al, 2012 (143)

### Setting

Households in 10 of the 16 'voivodeships' (provinces) in Poland. January 2003 to September 2010. Local data about TB burden in these communities not presented.

### Population

Seventy-eight patients with pulmonary tuberculosis, residing in 35 'Family-households' (either families 'linked by ties of consanguinity' or 'married couples'). Mostly Poles, although four Chechens. Ages 1-63 years.

### Details of strain-typing method used

The primary technique was 15 loci MIRU-VNTR, supplemented with spoligotyping.

### Definition of household

'…a group of two or more people residing in the same dwelling.'

### Definition of index case

'An 'index case' was defined as the first TB case diagnosed in the household, according to the date of collection of the first culture-positive sample. All subsequently identified cases in the same household were defined as secondary cases.'

### Definition of 'co-prevalent'

Not presented. However, the date of MTB isolation is presented for every case and all secondary case isolates were obtained in the same or subsequent year as the index case.

## Behr et al, 1998 (146)

### Setting

San Francisco County, 1 January 1991 to 31 December 1996. Local data about the TB burden in the county not presented.

### Population

The contacts of all people with infectious (pulmonary or laryngeal TB) and all people with TB under the age of fifteen.

### Details of strain-typing method used

IS 6110 based RFLP supplemented with polymorphic GC rich sequence based RFLP where the IS6110 pattern had five or fewer bands.

### Definition of household

'…those who shared the same front door with the index case…'

### Definition of index case

'the first case diagnosed…when the first case was a child (younger than 15 yr of age), the adult was considered the index case…'

### Definition of 'co-prevalent'

No maximum interval used. 45/54 secondary cases had disease at initial contact investigation. Of the 11 contacts that developed MTB subsequently, they did so after a median of 218 days.

## Bennett et al, 2002 (137)

### Setting

Five US states (Arkansas, Massachusetts, Maryland, Michigan, New Jersey) and selected sites in California and Texas, January 1996 to December 2000. Local data about the TB burden in these sites not presented.

### Population

Both adults and children.

### Details of strain-typing method used

IS 6110 based RFLP supplemented with Polymorphic GC-rich sequence typing or (later) spoligotyping where there were six or fewer IS 6110 bands(265).

### Definition of household

Not presented.

### Definition of index case

'…a source case was defined as a person with active TB identified in a routine contact investigation as the probable source of transmission to another person with active TB.'

### Definition of 'co-prevalent'

Not presented. However, 'secondary cases' were people 'with active TB identified in a routine contact investigation as having acquired TB from one source case.'

## Borrell et al, 2009 (141)

### Setting

Barcelona, Spain, between 1 January 2013 and 31 December 2014. Local 'incidence' (a notification rate, I believe) was 31.8 per 100,000 inhabitants in 2003 and 25.9 per 100,000 inhabitants in 2004.

### Population

All TB cases reported to the TB programme.

### Details of strain-typing method used

IS 6110 based RFLP supplemented with 12 loci MIRU-VNTR where six or fewer bands or where IS 6110 patterns differed by a single band.

### Definition of household

Not presented.

### Definition of index case

'…the patient who first manifested symptoms with a pulmonary localisation. When these data were not available or the patients were asymptomatic, the index case was considered to be…the patient who initiated treatment earliest.'

### Definition of 'co-prevalent'

Not presented.

## Buu et al, 2010 (118)

### Setting

Rural Vietnam, 1 January 2003 to 31 December 2006. Local data about the TB burden in the study site not presented.

### Population

Not clear but (I believe) the study recruited only adults with smear positive TB.

### Details of strain-typing method used

All isolates were typed by 15 loci MIRU-VNTR and spoligotyping. Some isolates were additionally typed by IS 6110 based RFLP.

### Definition of household

'…all persons who share the same floor and the same food…'

### Definition of index case

'All persons for whom TB had been diagnosed through December 31, 2004.'

### Definition of 'co-prevalent'

'…household members for whom TB was diagnosed within 24 months after enrolment of the index case-patient.'

## Glynn et al, 2015 (149)

### Setting

Karonga District, a rural area in Northern Malawi, 1997-2010. Data on the local TB burden not presented but readily available elsewhere(151).

### Population

Culture confirmed TB patients recruited using enhanced passive case finding.

### Details of strain-typing method used

Whole genome sequencing with transmission considered confirmed if ten or fewer SNP differences.

### Definition of household

Not defined.

### Definition of index case

In the data Judith Glynn shared with us, index cases needed to have lived in the same household as the secondary case 'whilst the index case was ill'.

### Definition of 'co-prevalent'

No maximum interval applied.

## Huh et al, 1995 (134)

### Setting

South Korea. The dates isolates were obtained is unclear and local TB burden data are not presented.

### Population

'Close contact' patients with pulmonary tuberculosis. The patients were aged 15-54 years.

### Details of strain-typing method used

RFLP using both IS 6110 and IS 1081.

### Definition of household

Not presented but Dr SJ Kim confirmed by email that all 'close contact' patients were living together.

### Definition of index case

Not presented and, actually, not required as in the one household with three cases all had the same RFLP type.

### Definition of 'co-prevalent'

Not presented.

## Inigo et al, 2003 (139)

### Setting

Three districts of Madrid, Spain, during 1997-1999. Local annual incidence 31.0, 29.2 and 30.2 per 100,000 in 2007, 2008 and 2009 respectively.

### Population

All culture positive TB patients.

### Details of strain-typing method used

IS 6110 based RFLP typing with spoligotyping on isolates with fewer than six bands.

### Definition of household

'…sharing of a mutual residence…'

### Definition of index case

Not presented.

### Definition of 'co-prevalent'

'We examined every case for contact with another TB patient in the two years prior to symptom onset…'

## Leung et al, 2013 (144)

### Setting

Hong Kong, 1997-2011. Local TB incidence declined from 109 per 100,000 population in 1997 to 74 per 100,000 population in 2009.

### Population

All MDR-TB cases notified between 1997 and 2006 plus their contacts.

### Details of strain-typing method used

IS 6110 based RFLP.

### Definition of household

'…persons living and sleeping in the same household as the index case for at least 1 month.'

### Definition of index case

These were the MDR-TB cases notified between 1997 and 2006.

### Definition of 'co-prevalent'

Reported both 'prevalent case[s]' who were 'detected in the initial screening' and 'secondary case[s]' who were detected in the subsequent surveillance period (to May 2011).

## Martin et al, 2009 (140)

### Setting

Nine urban districts in the south of Madrid, Spain, 2002-2006. No local TB burden estimates presented.

### Population

All residents diagnosed with culture positive TB between January 2002 and December 2006.

### Details of strain-typing method used

IS 6110 based RFLP with spoligotyping if five or fewer bands. 15 loci MIRU-VNTR on a subset of isolates

### Definition of household

Not presented.

### Definition of index case

Not defined.

### Definition of 'co-prevalent'

'…the existence of one or more secondary cases arising from the same source over a 1-year period after the index case was diagnosed.'

## Middelkoop et al, 2015 (148)

### Setting

A township near Cape Town with a TB notification rate of approximately 2000 cases per 100,000 people per year.

### Population

'All tuberculosis patients resident in the community and notified from 2001 to 2010.'

### Details of strain-typing method used

IS 6110 based RFLP with strains with fewer than 6 bands excluded from analysis.

### Definition of household

In the paper, most data was by residential plot, which might contain between 1 and 22 households. However, Dr Middelkoop provided me with data on residents of 13 shacks and I restricted my analysis to these individuals.

### Definition of index case

'…the first case diagnosed within the cluster.'

### Definition of 'co-prevalent'

'…subsequent patients in the clusters…'

## Sia et al, 2013 (145)

### Setting

A DOTS clinic in Manila in the Philippines, between 2001 and 2003. No data on the local TB burden presented.

### Population

Smear positive TB patients and their household contacts.

### Details of strain-typing method used

IS 6110 based RFLP, 12 loci MIRU VNTR and spoligotyping were performed on all isolates.

### Definition of household

Not presented.

### Definition of index case

I believe this was this was the first patient from the household diagnosed by the clinic. Index cases were smear positive and over the age of 18 years whereas secondary cases were not required to be smear positive and ranged in age from 10-60 years.

### Definition of 'co-prevalent'

Not presented but I believe these were individuals identified during household contact tracing.

## Verver et al, 2004 (77)

### Setting

Cape Town's Northern Suburbs, 1993-1998. Local TB notification rate 320 per 100,000 per year (bacteriologically confirmed disease only).

### Population

All diagnosed TB patients.

### Details of strain-typing method used

IS 6110 based RFLP. No secondary typing method used (266).

### Definition of household

'…a house and the associated informal dwelling or dwellings at the same address on the same plot of land.' (see Chapter 2 for details)

### Definition of index case

Not defined.

### Definition of 'co-prevalent'

Calculated as the total number of cases in the household minus one.

## Whalen et al, 2011 (147)

### Setting

Mulago Hospital, Kampala, Uganda. 1995-2004.

### Population

Cases of smear positive TB and their household contacts.

### Details of strain-typing method used

IS 6110 based RFLP with PGRS on samples with fewer than six RFLP bands.

### Definition of household

Not presented.

### Definition of index case

Smear positive

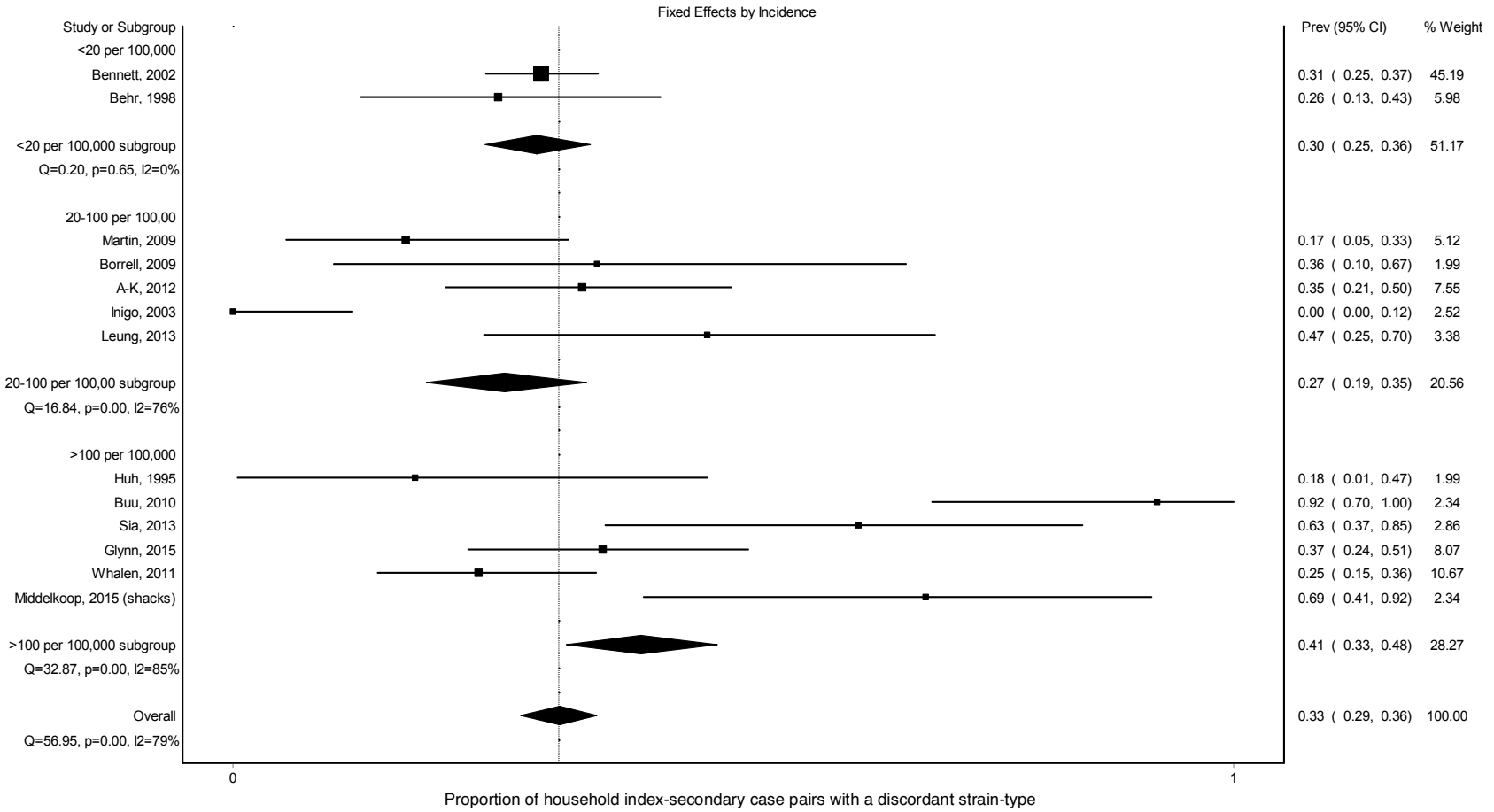
### Definition of 'co-prevalent'

I included both those Whalen termed 'co-prevalent' ('a tuberculosis case occurring within three months of the initial diagnosis in the incident case') and those he termed 'incident tuberculosis' which were cases occurring after three months. Contacts were followed for 2 years.

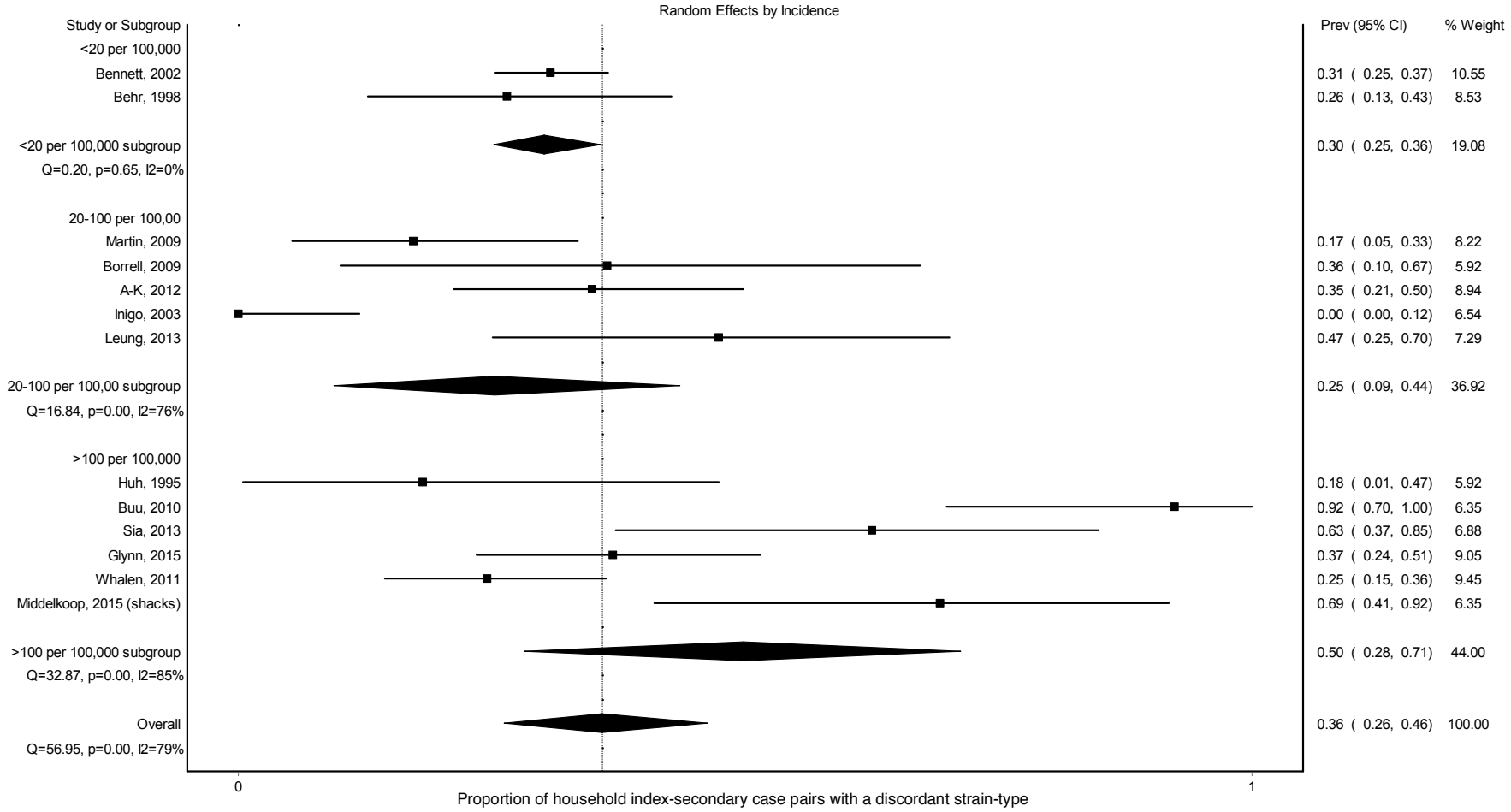
## Appendix 3 – Meta-analyses with Verver et al excluded

There was substantial heterogeneity in the unstratified meta-analysis ($I_2$ 77%).

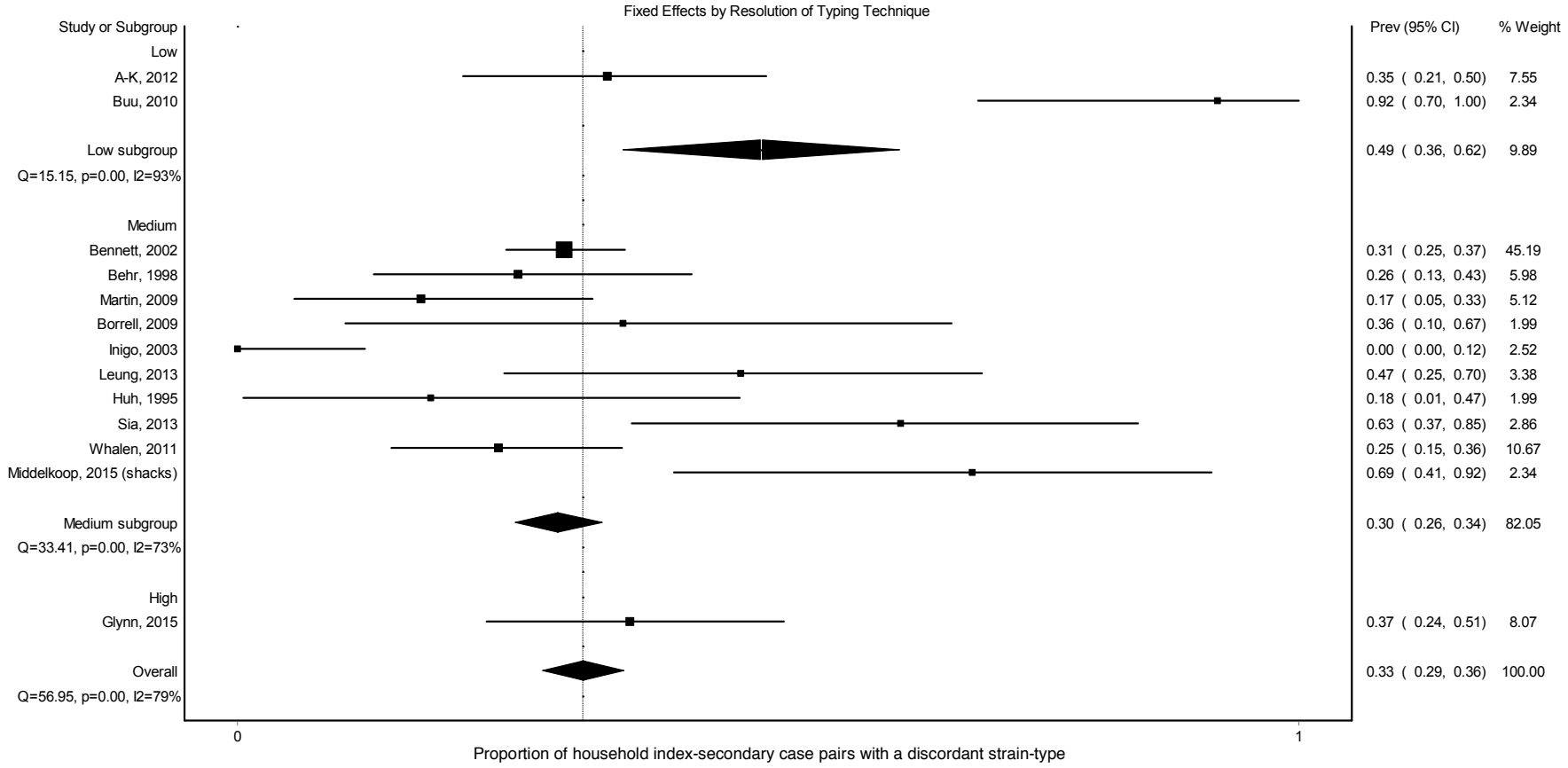
Supplementary figure 1. Fixed effects meta-analysis, stratified by national mid study TB incidence (excluding Verver et al).



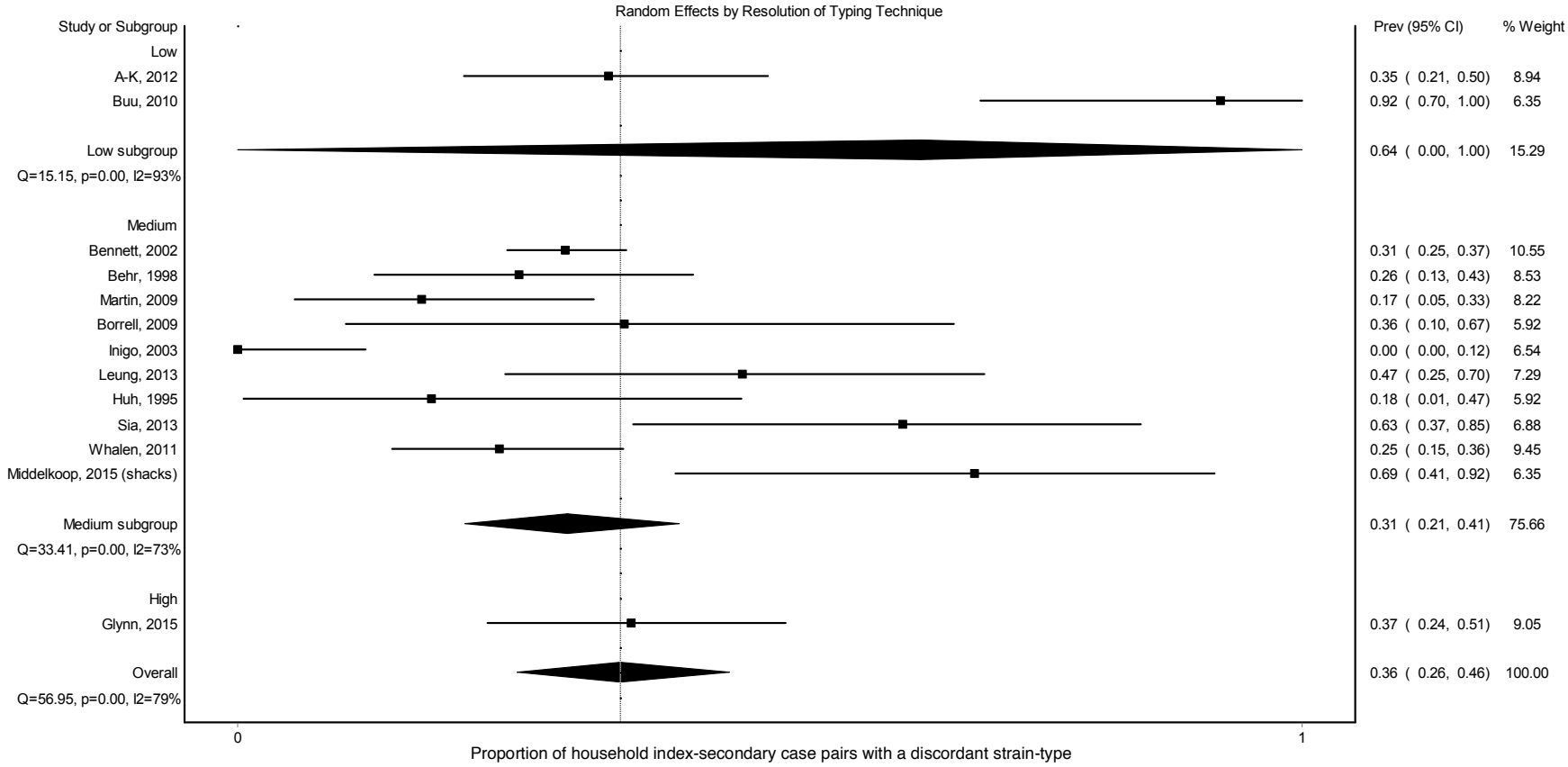
Supplementary figure 2. Random effects meta-analysis, stratified by national mid study TB incidence (excluding Verver et al).



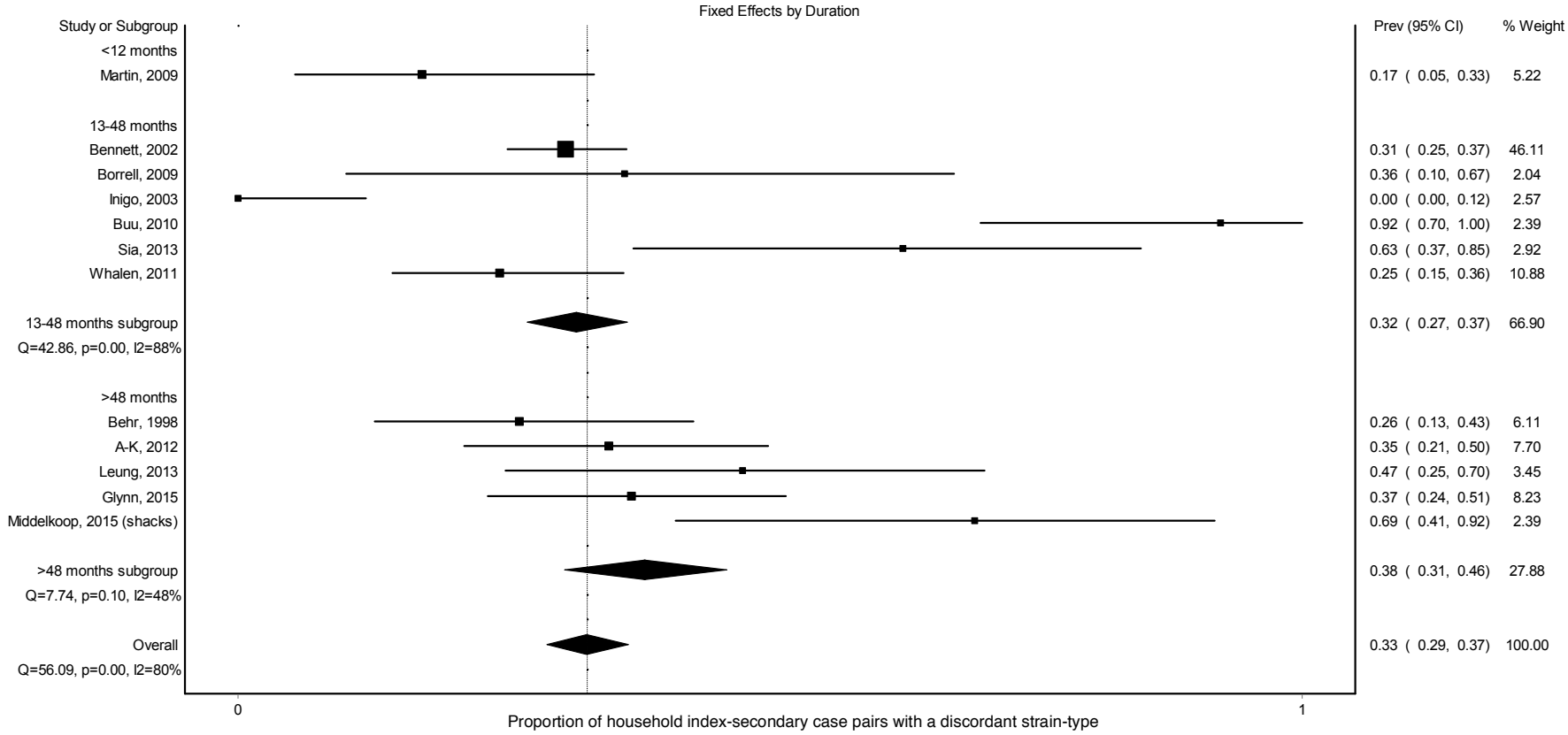
Supplementary figure 3. Fixed effects meta-analysis, stratified by the resolution of the strain typing technique (excluding Verver et al).



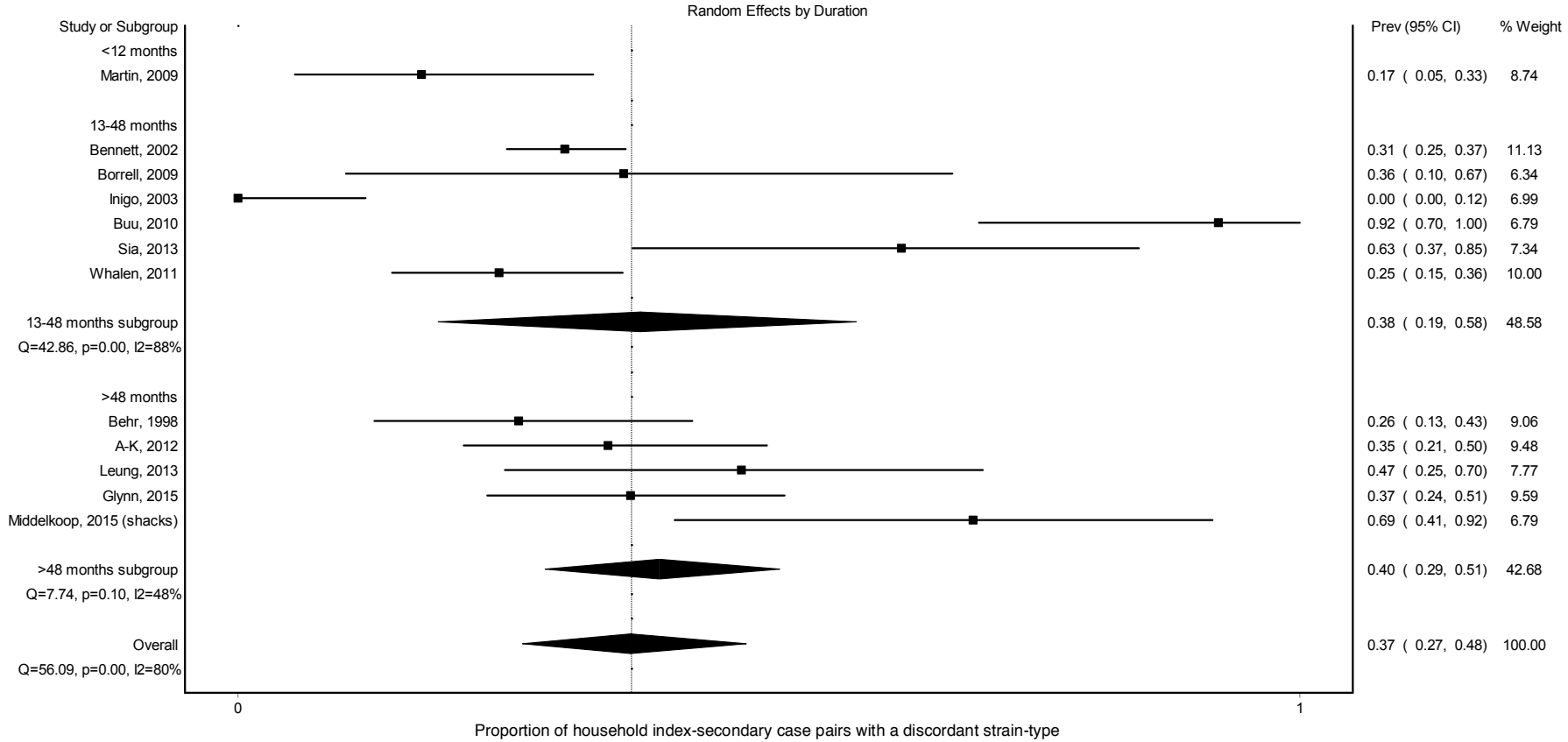
Supplementary figure 4. Random effects meta-analysis, stratified by the resolution of the strain typing technique (excluding Verver et al).



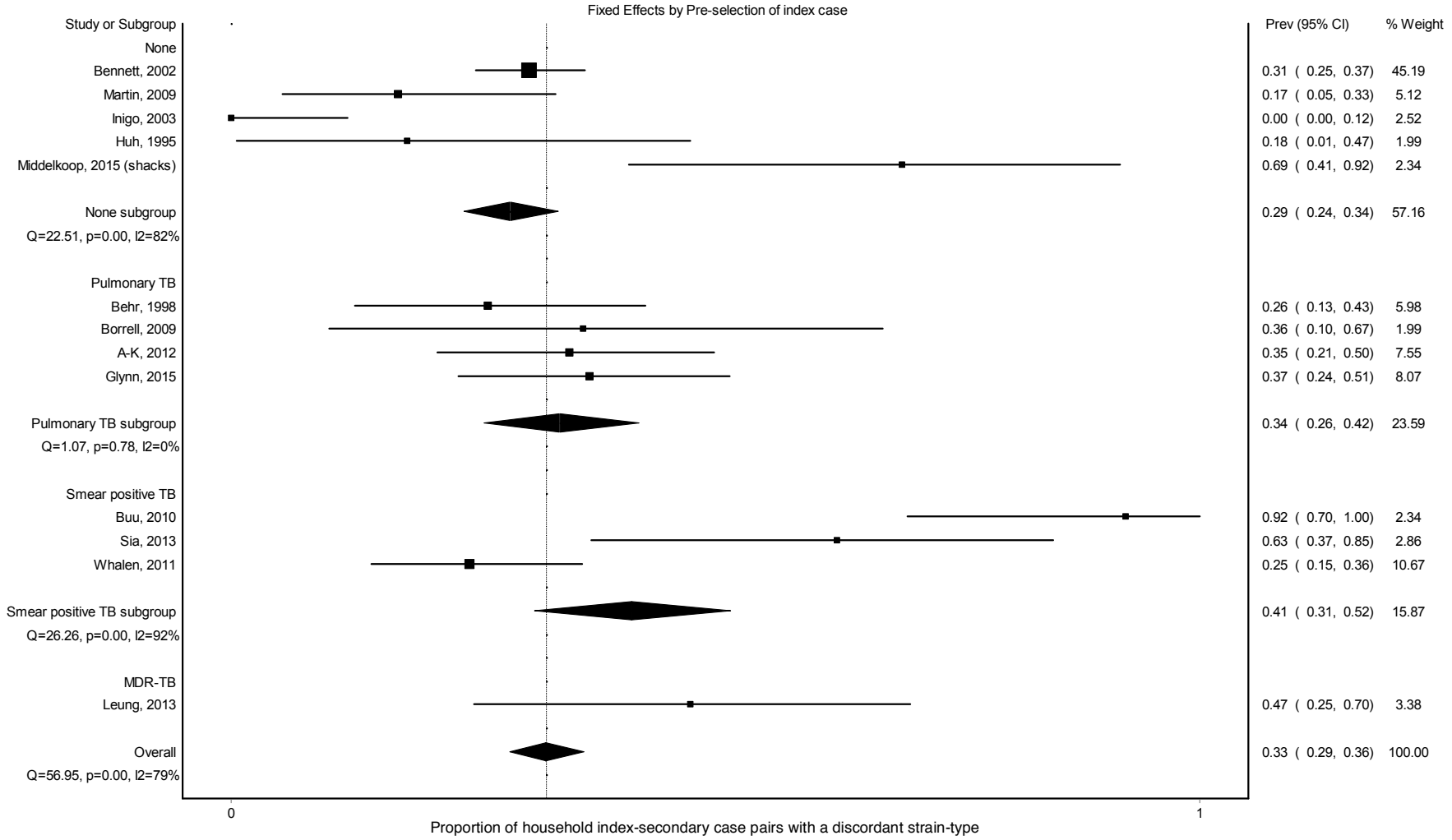
Supplementary figure 5. Fixed effects meta-analysis, stratified by the duration of sampling (excluding Verver et al).



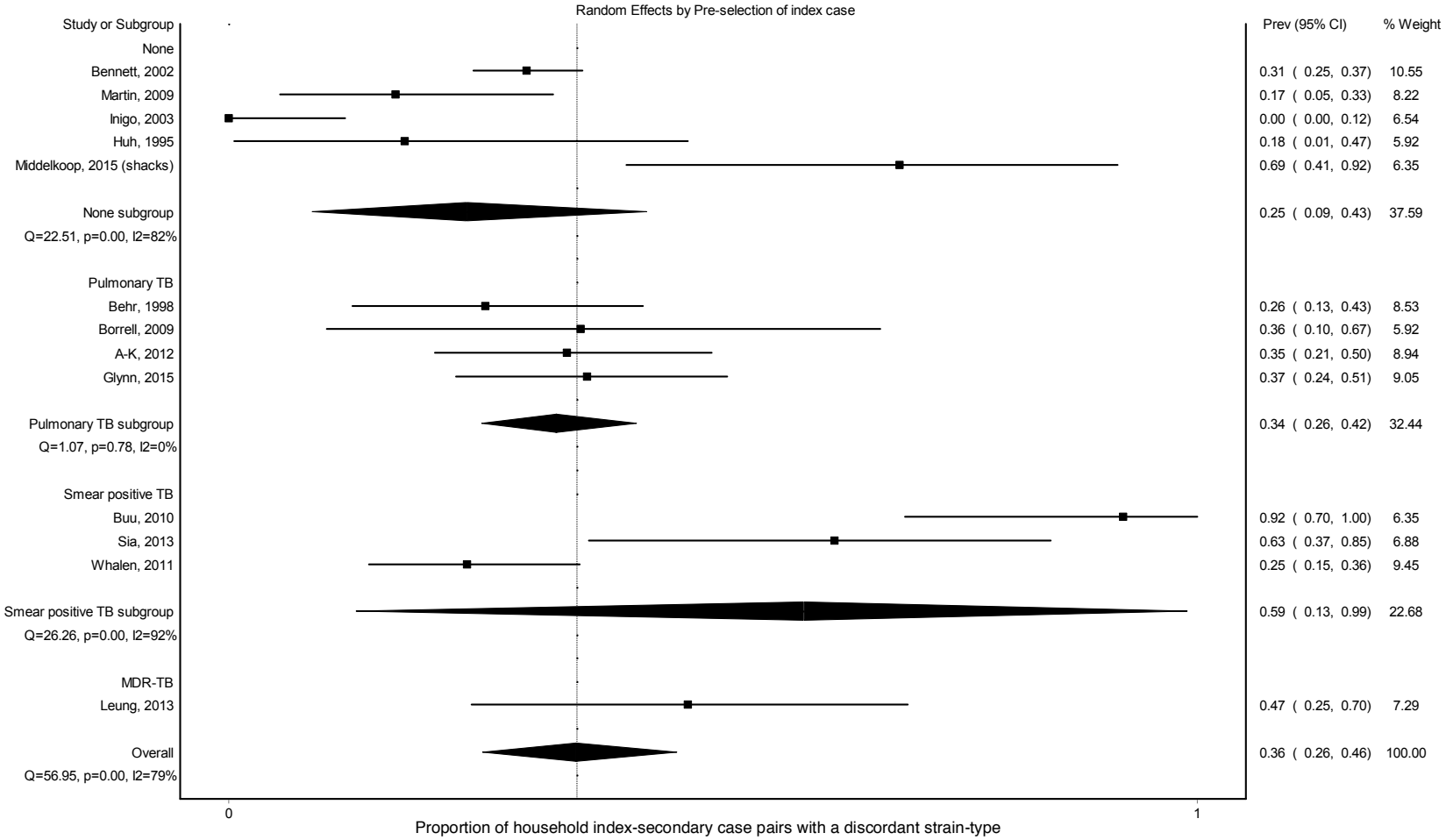
Supplementary figure 6. Random effects meta-analysis, stratified by the duration of sampling (excluding Verver et al).



Supplementary figure 7. Fixed effects meta-analysis, stratified by whether the index cases were pre-selected (excluding Verver et al).



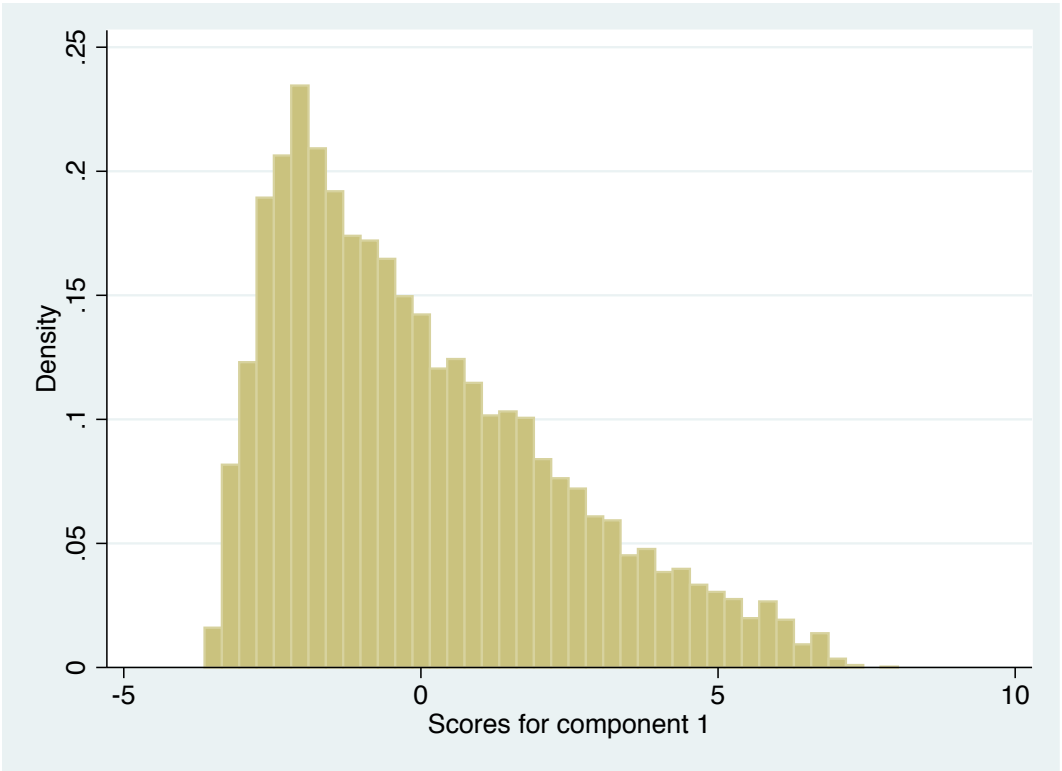
Supplementary figure 8. Random effects meta-analysis, stratified by whether the index cases were pre-selected (excluding Verver et al).



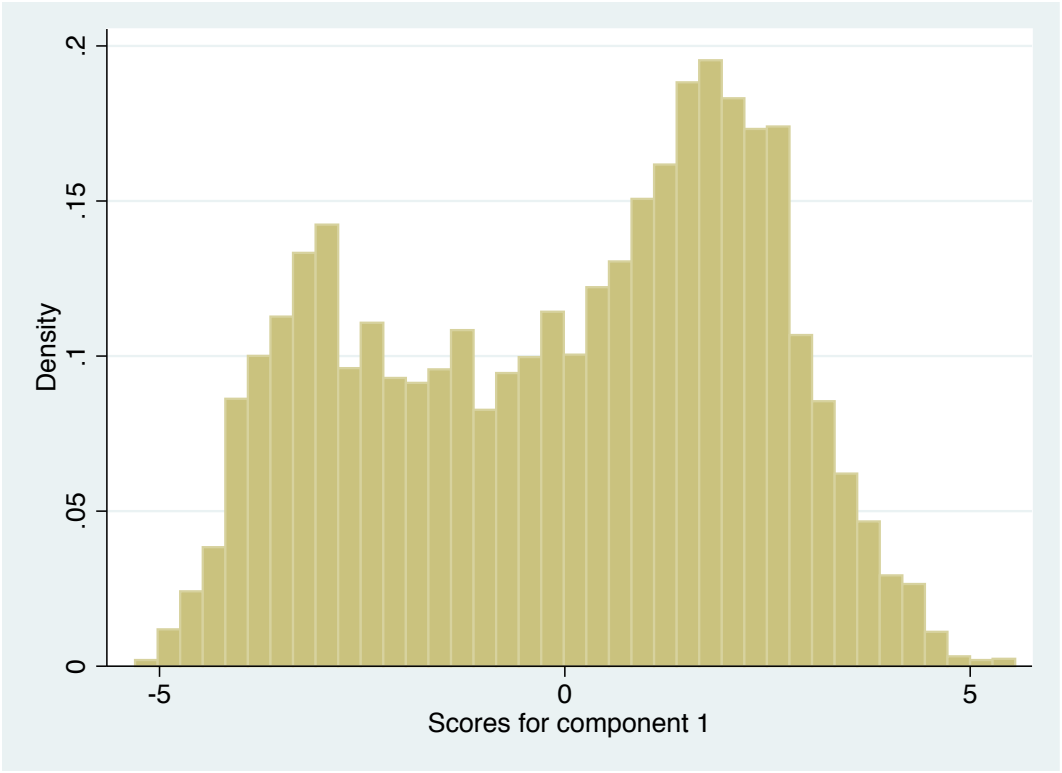
# Appendix 4 – The distribution of household wealth scores obtained using Principal Component Analysis, 2003-2013

These histograms plot the distribution of household wealth scores for households in the Africa Centre surveillance area over time. These scores were obtained using PCA on the same standard set of data on household ownership of assets that I used to construct the household wealth scores used in Chapters 3 and 4 of this thesis.

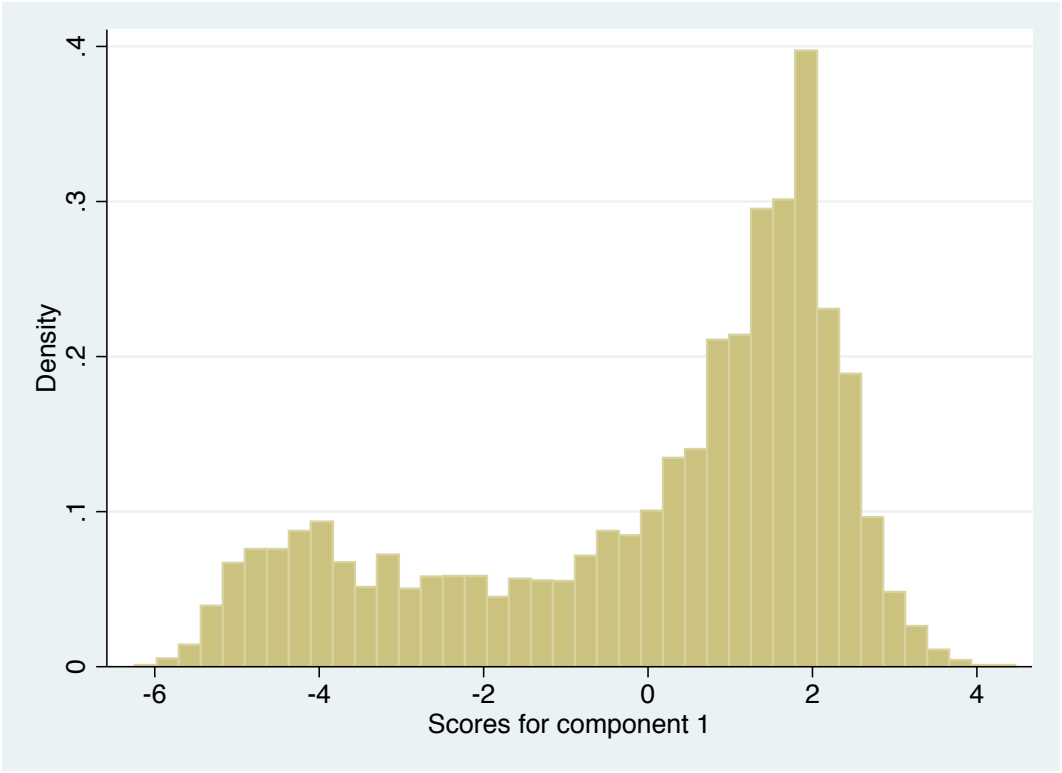
2003



2009


Density
0
.05
.1
.15
.2
-5
0
5
Scores for component 1

2013


Density
0
.1
.2
.3
.4
-6
-4
-2
0
2
4
Scores for component 1